Titre: Title:	Apprentissage de représentations pour la classification d'images biomédicales
Auteur: Author:	William Thong
Date:	2015
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Thong, W. (2015). Apprentissage de représentations pour la classification d'images biomédicales [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/1842/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/1842/
Directeurs de recherche: Advisors:	Samuel Kadoury, & Christopher J. Pal
Programme: Program:	Génie biomédical

UNIVERSITÉ DE MONTRÉAL

APPRENTISSAGE DE REPRÉSENTATIONS POUR LA CLASSIFICATION D'IMAGES BIOMÉDICALES

WILLIAM THONG
INSTITUT DE GÉNIE BIOMÉDICAL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE BIOMÉDICAL)
AOÛT 2015

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

APPRENTISSAGE DE REPRÉSENTATIONS POUR LA CLASSIFICATION D'IMAGES BIOMÉDICALES

présenté par : THONG William
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. BILODEAU Guillaume-Alexandre, Ph. D., président
M. KADOURY Samuel, Ph. D., membre et directeur de recherche
M. PAL Christopher J., Ph. D., membre et codirecteur de recherche
M. RAISON Maxime, Doctorat, membre

DÉDICACE

À ma famille.

REMERCIEMENTS

Je témoigne une sincère reconnaissance à mon directeur de recherche Samuel Kadoury et mon co-directeur de recherche Christopher J. Pal pour m'avoir accompagné et guidé tout au long de ma maîtrise. Leurs conseils, leur disponibilité, ou encore leur attention m'ont permis de compléter ce mémoire et d'acquérir une expérience inestimable.

Le programme de formation MÉDITIS a joué un rôle indéniable dans ma maîtrise par l'encadrement qu'ils m'ont offert. Une attention particulière revient à Nathalie Jourdain pour la coordination du programme et le temps passé à suivre chaque étudiant.

Je tiens également à remercier Guillaume-Alexandre Bilodeau et Maxime Raison pour le temps qu'ils ont accepté de m'accorder afin d'évaluer mon mémoire.

Le projet sur la colonne vertébrale n'aurait pas pu se réaliser sans Stefan Parent, Hubert Labelle et Carl-Éric Aubin. Je suis venu me greffer à un projet dont ils étaient les investigateurs principaux. Leur expertise a largement contribué aux travaux de ce mémoire. J'ai également une pensée pour les assistants de recherche du CHU Sainte-Justine, Nathalie, Christian et Soraya, sans qui rassembler autant de patients aurait été impossible.

Je souhaite remercier Nicolas Piché, à la base du projet sur la segmentation des reins, et avec qui les discussions ont toujours été très enrichissantes.

Je voudrais aussi remercier Julien Cohen-Adad, avec qui j'ai initialement commencé ma maîtrise. Alors étudiant au baccalauréat, j'ai réalisé de nombreux projets dans son laboratoire qui ont parfait mes connaissances en imagerie médicale.

Je ne saurais oublier J.P Lewis, qui m'a accueilli dans son laboratoire en Nouvelle-Zélande, et qui m'a montré l'envers du décor du monde du cinéma.

Une pensée chaleureuse va à tous les étudiants des cycles supérieurs que j'ai côtoyés à Polytechnique Montréal. Qu'ils soient des collègues de bureau, collègues au sein du département, ou désormais dans un nouveau laboratoire, nombreux sont devenus des amis proches.

Enfin, j'adresse mes remerciements les plus chers à ceux qui ont suivi de très près, voire de trop près, l'ensemble des tribulations survenues au cours de ma maîtrise, et qui se reconnaîtront sans aucune peine. Famille et amis m'ont offert un soutien sans faille qui a été d'une aide précieuse et indispensable.

RÉSUMÉ

La disponibilité croissante d'images médicales ouvre la porte à de nombreuses applications cliniques qui ont une incidence sur la prise en charge du patient. De nouveaux traits caractéristiques cliniquement pertinents peuvent alors être découverts pour expliquer, décrire et représenter une maladie. Les algorithmes traditionnels qui se basent sur des règles d'association manuellement construits font souvent défaut dans le domaine biomédical à cause de leur incapacité à capturer la forte variabilité au sein des données. L'apprentissage de représentations apprend plusieurs niveaux de représentations pour mieux capturer les facteurs de variation des données. L'hypothèse du projet de recherche du mémoire est que la classification par apprentissage de représentations apportera une information supplémentaire au médecin afin de l'aider dans son processus de décision. L'objectif principal, qui en découle, vise à étudier la faisabilité de l'apprentissage de représentations pour le milieu médical en vue de découvrir des structures cliniquement pertinentes au sein des données.

Dans un premier temps, un algorithme d'apprentissage non-supervisé extrait des traits caractéristiques discriminants des déformations de la colonne vertébrale de patients atteints de la scoliose idiopathique de l'adolescent qui nécessitent une intervention chirurgicale. Le sous-objectif consiste à proposer une alternative aux systèmes de classification existants qui décrivent les déformations seulement selon deux plans alors que la scoliose déforme le rachis dans les trois dimensions de l'espace. Une large base de données a été rassemblée, composée de 915 reconstructions de la colonne vertébrale issues de 663 patients. Des auto-encodeurs empilés apprennent une représentation latente de ces reconstructions. Cette représentation de plus faible dimension démêle les facteurs de variation. Des sous-groupes sont par la suite formés par un algorithme de k-moyennes++. Onze sous-groupes statistiquement significatifs sont alors proposés pour expliquer la répartition de la déformation de la colonne vertébrale.

Dans un second temps, un algorithme d'apprentissage supervisé extrait des traits caractéristiques discriminants au sein d'images médicales. Le sous-objectif consiste à classifier chaque voxel de l'image afin de produire une segmentation des reins. Une large base de données a été rassemblée, composée de 79 images tomographiques avec agent de contraste issues de 63 patients avec de nombreuses complications rénales. Un réseau à convolution est entrainé sur des patches de ces images pour apprendre des représentations discriminantes. Par la suite, des modifications sont appliquées à l'architecture, sans modifier les paramètres appris, pour produire les segmentations des reins. Les résultats obtenus permettent d'atteindre des scores élevés selon les métriques utilisées pour évaluer les segmentations en un court délai de calcul.

Des coefficients de Dice de 94,35% pour le rein gauche et 93,07% pour le rein droit ont été atteints.

Les résultats du mémoire offrent de nouvelles perspectives pour les pathologies abordées. L'application de l'apprentissage de représentations dans le domaine biomédical montre de nombreuses opportunités pour d'autres tâches à condition de rassembler une base de données d'une taille suffisante.

ABSTRACT

The growing accessibility of medical imaging provides new clinical applications for patient care. New clinically relevant features can now be discovered to understand, describe and represent a disease. Traditional algorithms based on hand-engineered features usually fail in biomedical applications because of their lack of ability to capture the high variability in the data. Representation learning, often called *deep learning*, tackles this challenge by learning multiple levels of representation. The hypothesis of this master's thesis is that representation learning for biomedical image classification will yield additional information for the physician in his decision-making process. Therefore, the main objective is to assess the feasibility of representation learning for two different biomedical applications in order to learn clinically relevant structures within the data.

First, a non-supervised learning algorithm extracts discriminant features to describe spine deformities that require a surgical intervention in patients with adolescent idiopathic scoliosis. The sub-objective is to propose an alternative to existing scoliosis classifications that only characterize spine deformities in 2D whereas a scoliotic is often deformed in 3D. 915 spine reconstructions from 663 patients were collected. Stacked auto-encoders learn a hidden representation of these reconstructions. This low-dimensional representation disentangles the main factors of variation in the geometrical appearance of spinal deformities. Sub-groups are clustered with the k-means++ algorithm. Eleven statistically significant sub-groups are extracted to explain how the different deformations of a scoliotic spine are distributed.

Secondly, a supervised learning algorithm extracts discriminant features in medical images. The sub-objective is to classify every voxel in the image in order to produce kidney segmentations. 79 contrast-enhanced CT scans from 63 patients with renal complications were collected. A convolutional network is trained on a patch-based training scheme. Simple modifications to the architecture of the network, without modifying the parameters, compute the kidney segmentations on the whole image in a small amount of time. Results show high scores on the metrics used to assess the segmentations. Dice scores are 94.35% for the left kidney and 93.07% for the right kidney.

The results show new perspectives for the diseases addressed in this master's thesis. Representation learning algorithms exhibit new opportunities for an application in other biomedical tasks as long as enough observations are available.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

Sigle	**Acronyme**	**Équivalent français**
AIS	Adolescent idiopathic scoliosis	Scoliose idiopathique de l'adolescent
ANN	Artificial neural networks	Réseaux de neurones artificiels
ASSD	Average symmetric surface distance	Distance moyenne symétrique surfacique
CHVA	Central hip vertical axis	Axe vertical centré sur les hanches
ConvNet	Convolutional network	Réseau à convolution
CSVL	Central sacral vertical line	Ligne verticale centrée sur le sacrum
CT	Computerized tomography	Tomodensitométrie
DC	Dice coefficient	Coefficient de Dice
FOV	Field-of-view	Champ de vision
HD	Hausdorff distance	Distance d'Haussdorff
MT	Main thoracic	Thoracique haute
PI	Pelvic incidence	Incidence pelvienne
PMC	Plane of maximum curvature	Plan de courbure maximale
ReLU	Rectified linear unit	Unité rectificatrice linéaire
SAE	Stacked auto-encoders	Auto-encodeurs empilés
TLL	Thoracolumbar/lumbar	Thoracolombaire/lombaire

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

L'apprentissage automatique se définit comme la capacité d'un agent à apprendre à prendre une décision à partir d'observations (Bengio et al., 2015). Dans le contexte biomédical, l'action de cet agent se traduit par une information additionnelle pour aider le médecin dans sa prise de décision. La prise en charge d'un patient se retrouve affectée à plusieurs étapes, que ce soit au niveau du diagnostic, du choix du traitement, du suivi dans le temps, ou encore dans l'intervention chirurgicale. Dans le cadre de ce mémoire, l'agent en question a pour rôle de classer des images biomédicales par apprentissage automatique en vue de découvrir des patrons de pathologies cliniquement pertinents. Ces opérations de classification demeurent à la base des outils d'aide à la décision médicale assistée par ordinateur.

Toutefois, la variabilité entre les patients pose de nombreux défis pour les algorithmes de classification traditionnels. Ces derniers ont pour la plupart été configurés et paramétrés sur de petits jeux de données ou sur une cohorte très spécifique. En d'autres mots, leur capacité de généralisation à de nouveaux patients demeure relativement faible. Au cours de la dernière décennie, l'apprentissage de représentations – un sous-domaine de l'apprentissage automatique – a connu un retour fulgurant, particulièrement dans le domaine de la vision par ordinateur et du traitement automatique du langage naturel. Ces algorithmes de représentations ont notamment permis de franchir un pas significatif en ce qui concerne la reconnaissance d'objets (Krizhevsky et al., 2012) et de la parole (Hinton et al., 2012). Les géants de l'Internet, que sont par exemple les GAFA (Google, Amazon, Facebook, Apple) ou encore Microsoft et Baidu, déploient aujourd'hui des solutions basées sur les algorithmes de représentations dans nombre de leurs produits. Leur utilisation fait désormais partie de notre quotidien sans que l'on ne le sache ou que l'on n'en ait forcément conscience.

La capacité à tirer profit de ces algorithmes de représentations reste néanmoins limité dans le domaine biomédical. À notre connaissance, peu d'articles de recherche dans le domaine biomédical se sont attelés à exploiter la capacité des algorithmes de représentations à apprendre sur de grands jeux de données et à généraliser leur performance sur de nouveaux cas. Les succès les plus marquants concernent notamment la segmentation de membranes dans des images de microscopie électronique (Ciresan et al., 2012), la segmentation de tumeurs cérébrales dans des images de résonance magnétique (Menze et al., 2014), ou encore la prédiction de la conformation de protéines (Xiong et al., 2015). La dimension éthique qui entoure ces données privées joue un rôle particulièrement prépondérant quant à la difficulté d'obtenir et de partager des données biomédicales. En effet, l'anonymisation des données et

des méta-données est cruciale pour assurer la protection de la vie privée du patient. Il est par exemple possible de reconnaitre une personne rien qu'en regardant les images médicales de son visage. À cela s'ajoutent l'interprétation et l'analyse de ces données qui requièrent une expertise médicale pointue difficile à accéder et à acquérir. L'ensemble fait en sorte que l'inertie autour de projets biomédicaux impliquant de l'apprentissage automatique est plus forte que d'autres projets au sein desquels ces barrières n'existent pas.

1.1 Définitions et concepts de base

La fonction principale des algorithmes d'apprentissage automatique consiste à apprendre des paramètres θ pour modéliser un jeu de données $\mathcal{D}$ contenant N exemples, et représentatif du problème que l'on cherche à résoudre. Chacun de ces algorithmes possède également des hyper-paramètres qui contrôlent leur capacité, c'est-à-dire leur performance à apprendre des paramètres pour modéliser $\mathcal{D}$ mais aussi à généraliser cet apprentissage sur un nouveau jeu de données. Chacun des exemples de $\mathcal{D}$ sera dénoté par $x^{(i)} \in \mathbb{R}^D$, où D correspond au nombre de dimensions de l'exemple et i est compris entre $[1, \cdots, N]$ exemples. Des cibles peuvent également être présentes pour étiqueter le jeu de données $\mathcal{D}$. Chaque exemple $x^{(i)}$ peut donc être relié à sa cible correspondante $y^{(i)}$, pour former $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$.

Il est à noter que selon le contexte, les termes *apprentissage de représentations*, *apprentissage profond*, ou encore *réseau de neurones artificiels* seront utilisés au sein de mémoire. Ces termes sont synonymes et correspondent à la même signification, à savoir des algorithmes dont l'architecture permet l'apprentissage de plusieurs niveaux de représentations.

1.1.1 Paradigmes d'apprentissage

Les algorithmes d'apprentissage automatique apprennent selon des objectifs particuliers qui peuvent être divisés en trois principales catégories : l'apprentissage supervisé, l'apprentissage non-supervisé et l'apprentissage par renforcement. Il existe également un paradigme d'apprentissage semi-supervisé qui mélange les apprentissages supervisé et non-supervisé.

Supervisé

L'apprentissage supervisé concerne les modèles prédictifs dont le rôle est d'apprendre à prédire les cibles pour des problèmes de régression ou de classification. Dans le cas d'un problème de régression, les cibles prennent des valeurs réelles, ainsi $y \in \mathbb{R}$. Par exemple, une tâche de prédiction pourrait être la prédiction du nombre de mois de survie d'un patient à partir de plusieurs métriques issues de tests cliniques. Dans le cas d'un problème de classification, les

cibles prennent des valeurs entières, plus communément appelées classes ou catégories. Si le problème de classification est binaire, alors $y \in \{0, 1\}$. Si le problème de classification est multiclasse, alors $y \in \{1, \cdots, L\}$ où L correspond au nombre de classes. Par exemple, une tâche de classification pourrait être la prédiction du stade de la maladie d'un patient à partir de plusieurs métriques issues de tests cliniques.

Non-supervisé

L'apprentissage non-supervisé concerne les modèles descriptifs où il n'existe pas de cible explicite. L'objectif est d'entrainer un modèle qui soit capable de déceler par lui-même des facteurs qui expliqueraient au mieux les données. En d'autres mots, l'idée est d'appendre une distribution de probabilité qui modélise le jeu de données. Plusieurs applications en découlent comme l'estimation de densité, la génération de nouveaux exemples, le partitionnement de données en sous-groupes, ou encore la réduction de dimensionnalité. Ces deux dernières tâches sont à la base des travaux des chapitres 4 et 5.

Semi-supervisé

L'apprentissage semi-supervisé concerne le cas où le jeu de données $\mathcal{D}$ est partiellement étiqueté. L'objectif est d'entrainer un modèle qui soit capable de tirer partie à la fois des cibles présentes mais aussi des données non étiquetés. Il est à noter que l'apprentissage semi-supervisé n'est pas abordé au cours de ce mémoire.

Renforcement

L'apprentissage par renforcement concerne l'apprentissage d'actions à effectuer dans un environnement changeant afin de maximiser une récompense totale. Il est à noter que l'apprentissage par renforcement n'est pas abordé au cours de ce mémoire.

1.1.2 Compromis biais-variance

L'apprentissage automatique consiste essentiellement à trouver les paramètres d'une fonction $\hat{f}_N$ à partir d'un jeu de données $\mathcal{D}$ contenant N exemples. Cette fonction $\hat{f}_N$ fait partie d'un ensemble de fonction F. Parmi cet ensemble de fonctions F, se trouve la fonction f_F^*, la meilleure fonction de l'ensemble, que l'algorithme d'apprentissage tente d'apprendre en estimant les paramètres θ de $\hat{f}_N$ à partir du jeu de données $\mathcal{D}$ disponible. Cette erreur d'estimation correspond à la *variance* du modèle. De plus, cet ensemble de fonctions F possède une capacité régulée par des hyper-paramètres, qui peut peut se définir comme la "richesse"

ou la “complexité” de F. Plus la capacité est grande, moins l’erreur d’approximation entre f_F^* et f^* (la fonction idéale pour la tâche à apprendre) sera petite. Cette erreur d’approximation correspond au *biais* du modèle. Ces principes sont illustrés à la Figure 1.1.

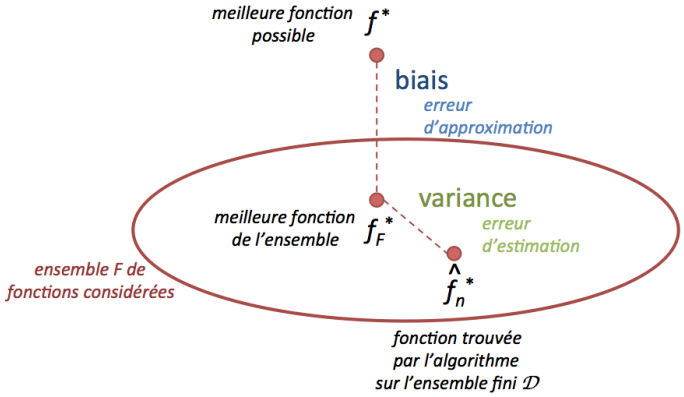


Figure 1.1 Compromis biais-variance.

Si toutefois la capacité du modèle est trop élevée, l’ensemble des fonctions F est trop riche pour décrire le jeu de données $\mathcal{D}$. Le modèle a en quelque sorte appris parfaitement à prédire les cibles du jeu de données, mais demeure incapable à généraliser sa performance sur un nouveau jeu de données. On parle alors de *sur-apprentissage*, où le modèle a un faible biais mais une variance trop élevée. Si au contraire la capacité du modèle est trop faible, l’ensemble des fonctions F est trop pauvre pour décrire le jeu de données $\mathcal{D}$. Le modèle prédit très mal les cibles du jeu de données car il est incapable de capturer la variabilité présente. On parle alors de *sous-apprentissage*, où le modèle a une faible variance mais un biais trop important. Une capacité optimale est alors nécessaire, c’est-à-dire un compromis entre le biais et la variance, pour permettre une meilleure généralisation sur un nouveau jeu de données comme l’illustre la Figure 1.2

L’heuristique principale pour éviter ces phénomènes de sur- et sous-apprentissage consiste à diviser un jeu de données en plusieurs partitions. Cela peut se traduire simplement en trois différentes partitions : $\mathcal{D}_{entrainement}$, $\mathcal{D}_{validation}$, $\mathcal{D}_{test}$. À des fins de simplicité, nous nous référerons à ces partitions respectivement par $\mathcal{D}_{train}$, $\mathcal{D}_{valid}$, $\mathcal{D}_{test}$. $\mathcal{D}_{train}$ représente le jeu de données sur lequel le modèle est entrainé pour qu’il ajuste ses paramètres. $\mathcal{D}_{valid}$ représente le jeu de données sur lequel se fait la sélection des hyper-paramètres et l’évaluation de la performance du modèle. Quant à $\mathcal{D}_{test}$, il consiste en un jeu de données pour évaluer la performance de généralisation finale. Le modèle ne doit en aucun cas voir les données de $\mathcal{D}_{test}$ pour que son évaluation ne soit pas biaisée. Lorsque le nombre d’exemples inclus dans $\mathcal{D}$ n’est pas assez important, une autre solution consiste à réaliser une validation croisée qui

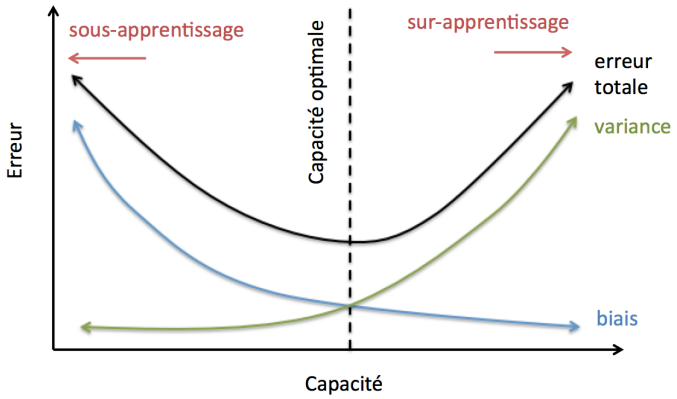


Figure 1.2 Capacité optimale pour éviter un sous- ou sur-apprentissage.

procède à une division en k-plis de $\mathcal{D}$. Le jeu de données initial $\mathcal{D}$ est divisé en k partitions, $k-1$ partitions sont sélectionnées pour l'entrainement et une partition pour la validation. Par exemple, une validation croisée en 5-plis divise $\mathcal{D}$ en cinq partitions avec une répartition des exemples dans chacun des sous-jeux de données réalisée aléatoirement. Cinq différentes évaluations sont alors réalisées pour couvrir toutes les possibilités. La performance finale du modèle est déterminée par la moyenne des erreurs de chacun des jeux de validation. Au cours de ce mémoire, seule la division en trois différentes partitions ($\mathcal{D}_{train}$, $\mathcal{D}_{valid}$, $\mathcal{D}_{test}$) sera utilisée pour rechercher les meilleurs hyper-paramètres et paramètres du modèle. Il est à noter qu'une fois tous les hyper-paramètres fixés, le modèle peut également être raffiné avec un jeu de données combinant $\mathcal{D}_{train}$ et $\mathcal{D}_{valid}$.

1.2 Éléments de la problématique et objectifs de recherche

L'expansion et la démocratisation des appareils d'acquisition d'imagerie médicale produit aujourd'hui une quantité incroyable de données liées au patient. Avoir accès à cette mine d'information ouvre la porte à l'application de l'apprentissage automatique pour forer les images médicales. Les algorithmes traditionnels font parfois défaut lors de la généralisation à de nouveaux cas à cause notamment de la forte variabilité inter- et intra-sujet. Les algorithmes d'apprentissage profonds s'attaquent à ce problème en apprenant plusieurs niveaux de représentations, capables de capturer la forte et riche variabilité des données contrairement aux heuristiques basées sur des règles d'association manuellement construites.

L'hypothèse sous-jacente du mémoire est que la classification par apprentissage de représentations apportera une information supplémentaire à haute valeur ajoutée pour le médecin qui aurait été difficile à obtenir avec des algorithmes basés sur des règles d'association tradi-

tionnelles. Par conséquent, l'objectif principal du projet de recherche s'attèle à l'étude de la faisabilité de l'apprentissage de représentations pour le milieu médical afin d'encourager la découverte de structures cliniquement pertinentes présentes dans les données. Une approche versatile a été adoptée pour la classification de données médicales par des algorithmes d'apprentissage de représentations de manière non-supervisée et supervisée. Plus spécifiquement, ce mémoire s'articulera autour de deux sous-objectifs qui abordent deux thèmes biomédicaux différents que sont la classification d'une cohorte en sous-groupes cliniquement significatifs et la classification de voxels pour la segmentation d'organes dans des images médicales.

Dans un premier temps, l'objectif est de découvrir des sous-groupes au sein d'une cohorte de patients atteints de la scoliose idiopathique de l'adolescent (AIS) et démontrant des déformations aux niveaux thoracique et lombaire pour proposer une alternative à la classification existante des déformations scoliotiques. La scoliose se définit comme une déformation complexe de la colonne vertébrale dans les trois dimensions de l'espace, résultant en des déformations structurelle, latérale et rotative de la colonne vertébrale. Seule 1 à 3 % de la population à risque, c'est-à-dire les enfants entre 10 et 16 ans, est touchée par l'AIS qui se caractérise par un angle de Cobb supérieur à 10° (Weinstein et al., 2008). Cette maladie dont l'étiopathogénie est inconnue apparaît autour de la puberté. Cependant, seule 0.25% de cette population à risque nécessitera un traitement suite la progression de la courbure de leur colonne vertébrale (Asher and Burton, 2006). Les traitements non-opératoires ont pour but de prévenir la progression de la courbure de la colonne vertébrale. Les approches varient toutefois, et leur impact sur la réduction de la courbure demeure encore à être évaluée pour éviter des études controversées (Asher and Burton, 2006; Weinstein et al., 2008). Par exemple, l'Amérique du Nord aura tendance à privilégier des traitements par corset alors que l'Europe privilégiera les traitements par kinésithérapie (Weinstein et al., 2008). Les traitements opératoires ont pour but de corriger la courbure dans les trois dimensions de manière permanente tout en limitant les complications à court et long terme. Les cas nécessitant une intervention chirurgicale sont notamment l'intérêt de ce mémoire. Une opération chirurgicale est souvent nécessaire lorsque l'angle de Cobb de la courbure principale dépasse 40°. Cette intervention consiste à effectuer une arthrodèse, c'est-à-dire une fusion de certaines vertèbres avec par exemple des tiges métalliques. La classification de ces patients est primordiale pour à la fois comprendre et caractériser la scoliose, mais aussi pour guider le chirurgien orthopédique dans les recommandations de traitements du patient et son le suivi à long terme. Toutefois, les systèmes de classification actuels ne tiennent en compte que deux dimensions de l'espace (King et al., 1983; Lenke et al., 2001), c'est-à-dire qu'ils n'exploitent pas complètement les informations sur la courbure de la colonne vertébrale. L'apprentissage de représentations a alors pour objectif d'extraire des traits caractéristiques discriminants de ces courbures, qui nécessitent

une intervention chirurgicale, pour proposer une alternative aux systèmes de classification existants.

Dans un second temps, l'objectif est de détecter, localiser et segmenter les reins dans des images tomodensitométriques issues d'une autre cohorte de patients atteints d'anévrisme de l'aorte abdominale. La segmentation d'organes au sein d'images médicales est à la base de nombreuses applications dans le domaine des diagnostics assistés par ordinateur, du traitement en radiothérapie, du guidage en radiologie interventionnelle ou encore du suivi du patient. La segmentation d'organes se définit comme un outil pour délimiter les régions d'intérêt pour le médecin. Cette délimitation peut alors servir de guide pour les instruments médicaux dans le cadre de radiothérapie où seule la zone segmentée est traitée, ou dans le cadre de la radiologie interventionnelle pour guider le radiologue dans ses gestes. Des métriques quantitatives peuvent également être calculées pour mesurer la progression dans le temps des déformations observées. Cependant, la segmentation de zones d'intérêt dans les images médicales est souvent considérée comme longue et fastidieuse. D'autant plus que cette tâche requiert parfois de travailler au voxel près pour obtenir une délimitation cliniquement correcte. Des méthodes de segmentation automatiques, robustes, exactes et précises sont alors nécessaires pour résoudre les défis posés par les applications cliniques. L'apprentissage de représentations a alors pour objectif d'extraire des traits caractéristiques discriminants au sein des images médicales pour segmenter les reins chez des patients présentant de nombreuses complications rénales.

Ce mémoire a donc pour objectif d'évaluer l'application des algorithmes d'apprentissage de représentations sur deux problèmes médicaux concrets. Les travaux de ce mémoire se concentreront d'abord sur une méthode d'apprentissage non-supervisée pour découvrir de nouveaux sous-groupes au sein d'une cohorte de patients atteints de l'AIS ; puis sur une méthode d'apprentissage supervisée pour la segmentation de reins dans des images médicales.

1.3 Plan du mémoire

Le mémoire présente dans un premier temps au sein du chapitre 2 une revue de littérature qui aborde les principes d'anatomie et de physique médicale avant d'exposer l'état de l'art concernant la construction de réseaux de neurones profonds. Le chapitre 3 présente la méthodologie abordée pour répondre aux deux objectifs spécifiques établis pour le projet de recherche. Les chapitres 4, 5 et 6 concernent les publications scientifiques produites au cours de ce projet de recherche. Le chapitre 7 présente des résultats complémentaires qui n'ont pu être ajoutés dans l'article du chapitre 6. Le chapitre 8 discute du projet de manière générale tandis que le dernier chapitre conclut ce mémoire en revenant sur les éléments clés.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce chapitre relate une revue de littérature des concepts abordés au cours de ce mémoire. La section 2.1 présente les notions d'anatomie, utiles pour saisir la portée médicale des chapitres 4, 5 et 6, ainsi que la terminologie couramment utilisée pour l'analyse de patients atteints d'AIS. La section 2.2 aborde les fondements de l'imagerie médicale à base de photons en couvrant notamment la radiographie traditionnelle de projection (section 2.2.1), le système récent EOS (section 2.2.2) et la tomodensitométrie (section 2.2.3). Les images médicales utilisées au cours des chapitres 4, 5 et 6 sont issues de ces modalités d'imagerie. Finalement, les sections 2.4, 2.5 et 2.6 montrent comment construire un réseau de neurones artificiels ainsi que les méthodes actuellement utilisées pour son entrainement.

2.1 Anatomie et terminologie

2.1.1 Colonne vertébrale

La colonne vertébrale, parfois appelée rachis, est constituée d'une succession de vertèbres, séparées entre elles par des disques intervétébraux qui ont pour rôle de former l'armature du tronc et de protéger le système nerveux central (Baqué and Maes, 2008).

Figure 2.1 Illustratientraion de la colonne vertébrale et de ses différentes régions selon les plans coronal (antérieur), coronal (postérieur) et sagittal, tirée de Wikimedia Commons (2010).

La colonne vertébrale se définit comme une superposition de cinq grandes zones : 7 vertèbres cervicales (C1–C7) ; 12 vertèbres thoraciques (T1–T12) ; 5 vertèbres lombaires (L1–L5) ; le sacrum formé de 5 vertèbres sacrées fusionnées ; et le coccyx formé de 4 ou 5 vertèbres coccygiennes fusionnées. Dans le plan coronal, la colonne vertébrale est médiane (c'est-à-dire au centre du corps humain) et verticale. Par contre, dans le plan sagittal, elle décrit une courbe composée d'une alternance de concavité dorsale (lordose) et de concavité ventrale (cyphose). Plus familièrement, on parle également d'une alternance de creux et de bosse dans le dos.

La scoliose idiopathique de l'adolescent entraine une perturbation au niveau de ces courbures, et ce, dans les trois dimensions de l'espace. Dans le plan coronal, la colonne vertébrale a tendance à décrire une forme de esse (d'où le nom de scoliose). Dans le plan sagittal, cela se manifeste notamment par des hypercyphoses et/ou des hyperlordoses, voire des hypocyphoses et/ou des hypolordoses. Dans le plan axial, un mouvement de rotation entre les vertèbres se crée. Dans le cadre de ce mémoire, nous ne considèrerons que les déformations des régions thoracique et lombaire qui concernent quatre grands types de scoliose : la scoliose thoracique qui se décrit par une déformation de la colonne vertébrale au niveau du thorax, créant alors une bosse dans le dos appelée gibbosité ; la scoliose thoraco-lombaire qui concerne une déformation plus longue qui touche également le segment lombaire ; la scoliose lombaire qui se caractérise par une déformation au niveau de la taille avec une gibbosité peu prononcée ; et enfin, la scoliose double qui présente deux courbures, le plus souvent une thoracique droite et une lombaire gauche.

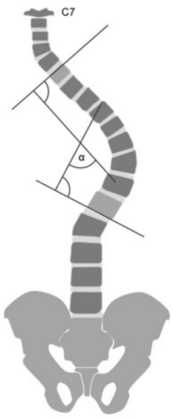


Figure 2.2 Calcul de l'angle de Cobb, image tirée de Waldt et al. (2014).

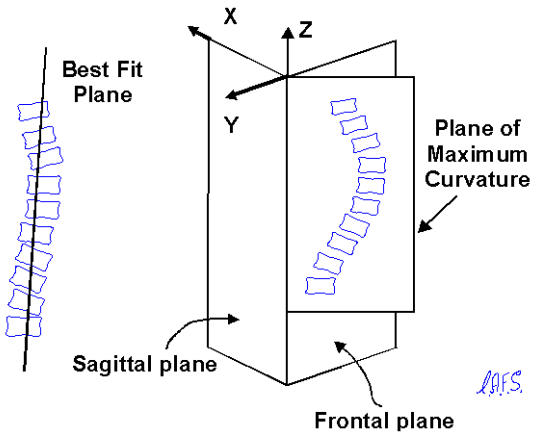


Figure 2.3 Plan de courbure maximale, image tirée de Stokes (1994).

Angle de Cobb

L'angle de Cobb [en degré] est la métrique de référence pour l'évaluation de la scoliose définie par la Scoliosis Research Society (Stokes, 1994). L'angle de Cobb d'une courbure correspond à l'angle entre la droite tangente au plateau supérieure de la vertèbre céphalade (c'est-à-dire du haut de la courbure) et la droite tangente au plateau inférieur de la vertèbre caudale (c'est-à-dire du bas de la courbure) (Waldt et al., 2014). La Figure 2.2 montre un exemple de calcul pour une courbure thoraco-lombaire.

Autres mesures pour quantifier la scoliose

Des mesures régionales, c'est-à-dire au niveau d'une courbure, peuvent complémenter l'angle de Cobb. Le plan de courbure maximale (PMC) est le plan vertical, illustré à la Figure 2.3, selon lequel l'angle de Cobb est le plus important (Stokes, 1994). L'idée est de projeter la colonne vertébrale sur un plan vertical. Ce plan vertical tourne autour de l'axe z (qui passe par la partie inférieure de la vertèbre caudale) jusqu'à ce que l'on obtienne l'angle de Cobb le plus important.

La vertèbre apicale est la vertèbre la plus déformée au sein d'une courbure. La mesure de sa rotation axiale [en degré] constitue une métrique importante pour le chirurgien lors de la correction de la courbure durant l'intervention chirurgicale.

L'incidence pelvienne [en degré] se définit comme l'angle entre la droite qui connecte les têtes fémorales au milieu du plateau du sacrum et la droite perpendiculaire à ce point (Legaye et al., 1998). Cette métrique est constante et unique chez chaque humain et demeure corrélée à la lordose lombaire.

La déviation latérale du tronc [en degré] sur le plan coronal indique la balance générale de la colonne vertébrale (Waldt et al., 2014). Sa mesure consiste en la distance entre deux droites parallèles, la droite passant par le centre de C7 et la droite passant par le centre du premier segment sacré, et permet de déterminer si la colonne vertébrale penche vers la gauche (valeur négative) ou vers la droite (valeur positive).

Représentation daVinci

La représentation daVinci, illustrée à la droite de la Figure 2.4 est une méthode de visualisation pour aider l'interprétation des indices calculés dans le PMC pour les trois segments d'intérêt (thoracique, thoracolombaire et lombaire). Le centre de chaque représentation correspond à l'axe partant du centre du premier segment du sacrum (Sangole et al., 2009). L'axe

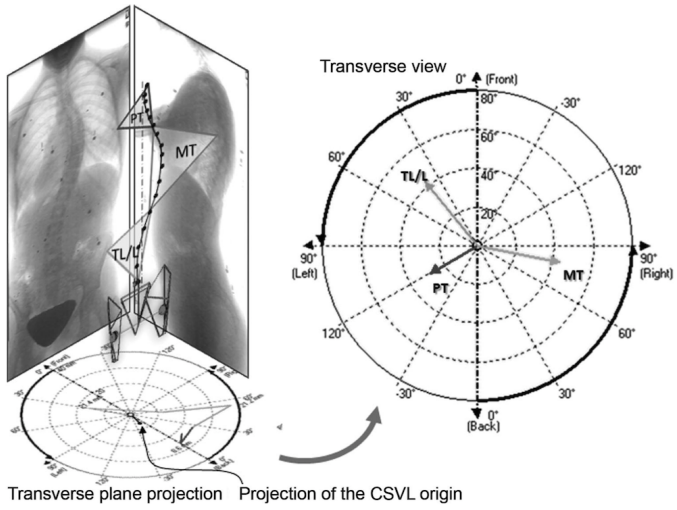


Figure 2.4 Représentation daVinci introduite par Sangole et al. (2009).

horizontal reflète l'amplitude de la déformation sur le plan coronal tandis que l'axe vertical reflète l'amplitude de la déformation sur le plan sagittal. Trois flèches partent du centre pour décrire les déformations dans les trois segments. La longueur d'une flèche est proportionnelle à la déformation de la courbure mesurée par l'angle de Cobb tandis que l'orientation d'une flèche correspond à l'angle de rotation du PMC.

2.1.2 Reins

Le rôle majeur des reins est de filtrer le sang pour éliminer les toxines et assurer l'équilibre homéostatique, ce qui engendre l'excrétion d'urine. Le rein prend la forme d'un haricot avec une hauteur de 12 cm, une largeur de 6 cm et une épaisseur de 3 cm, le tout pour un poids compris entre 130 et 140 g (Baqué and Maes, 2008). Les reins ont également une fonction endocrine. En effet, sur chacun des reins se trouve une glande surrénale dont la partie périphérique est responsable de la sécrétion de corticoïdes et la partie centrale des catécholamines.

Le rein est entouré d'une capsule fibreuse qui le protège. Sa structure est composée d'une partie périphérique, appelée cortex rénal, qui comporte les glomérules chargés de sang à filtrer ainsi que d'une partie centrale, appelée médulla, formée de pyramides et colonnes rénales. Sa vascularisation se fait par l'artère rénale provenant de l'aorte abdominale et par la veine rénale repartant vers la veine cave inférieure.

L'unité de base du rein est le néphron qui s'occupe de la filtration du sang. Un rein contient environ un million de néphrons (Ramé and Thérond, 2012). Une pyramide rénale est formée

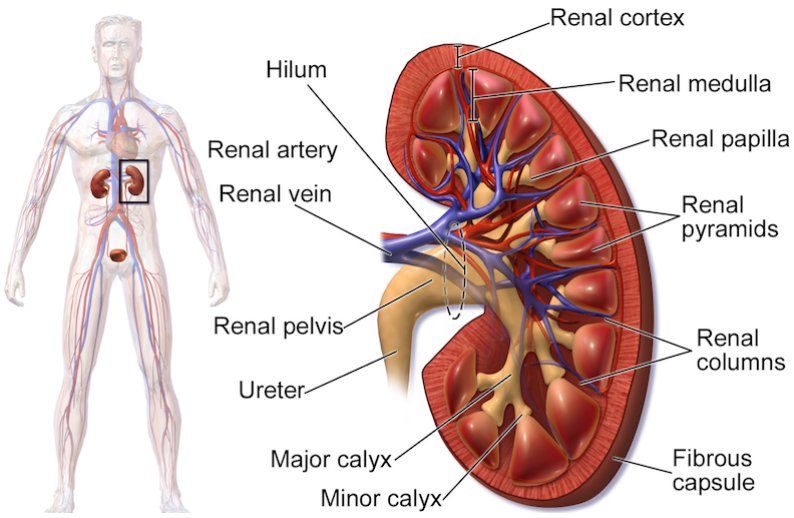


Figure 2.5 Illustration du rein et description des structures internes, tirée de Wikiversity Journal of Medicine (2014).

de plusieurs néphrons qui filtrent le sang provenant des vaisseaux présents sur une colonne rénale. Les résidus excrétés par les papilles rénales, c'est-à-dire l'extrémité d'une pyramide, se déversent dans un petit calice. L'urine parcourt par la suite le grand calice, qui constitue un ensemble de petits calices, pour finalement se décharger dans l'urètre par l'hile du rein.

Dans le cadre de ce mémoire, seule la définition macroscopique du rein sera considérée. Le rein est donc délimité par la capsule fibreuse ainsi que par l'hile du rein. Nous ne nous attarderons pas à segmenter les parties intérieures du reins. La base de données rassemblée au chapitre 6 concerne des patients atteints d'anévrismes de l'aorte abdominale ayant subi la pose d'un stent par chirurgie. Cette opération s'accompagne le plus souvent de complications sévères au niveau des reins (Godet et al., 1997). Les patients de cette base de données présentent donc pour la plupart des reins dysfonctionnels, ce qui complexifie la tâche de segmentation étant donnée la forte variabilité due aux différentes déformations engendrées par les pathologies.

2.2 Imagerie par rayons X

2.2.1 Rayons X

Les rayons X sont une forme de rayonnement électromagnétique dont la longueur d'onde est comprise entre 0,01 et 10 nm (Fanet, 2010). Un rayon X correspond au rayonnement d'un électron lors d'une transition entre deux niveaux d'un atome qui provient d'une accélération par une haute tension électrique. Ils ont la propriété de pouvoir traverser les tissus humains,

ce qui a fait d'eux une méthode populaire pour l'imagerie médicale. Cette propriété a été découverte par Wilhelm Röntgen en 1895, ce qui lui a valu le Prix Nobel de Physique en 1901. Les rayons X présentent toutefois un effet néfaste sur l'ADN, en cassant les liaisons des acides nucléiques. Des erreurs de retranscription de l'ADN surviennent alors, et engendrent possiblement des mutations génétiques pouvant conduire à des cancers (Fanet, 2010).

Le phénomène physique exploité par l'imagerie médicale est l'absorption des rayons X par la matière, en l'occurrence le corps humain, qu'ils traversent. Ce phénomène obéit à la loi de Beer-Lambert qui permet d'estimer l'intensité d'un signal I après qu'il a traversé un corps d'épaisseur l et caractérisé par un coefficient d'absorption linéaire μ :

$$I = I_0 \exp(-\mu l) \tag{2.1}$$

où I_0 représente l'intensité du signal d'entrée, c'est-à-dire le signal émis par la source de rayons X. Le coefficient d'absorption linéaire est plus faible pour les matériaux au faible numéro atomique et plus élevé pour les matériaux au numéro atomique élevé (Fanet, 2010). Ainsi, les rayons X sont très peu absorbés par les tissus mous alors qu'ils le sont beaucoup plus pour les structures osseuses, ce qui se répercute sur la qualité de l'image qui a un meilleur contraste de l'image pour les os. Cette propriété fait en sorte que les rayons X sont notamment très utilisés pour imager les structures osseuses tandis que les autres modalités d'imagerie, telles que l'imagerie par résonance magnétique, sont privilégiées pour imager les tissus mous.

L'instrumentation de la radiographie par rayons X comporte essentiellement deux composantes principales : une source responsable de l'émission des rayons et un détecteur qui capte le flux de rayons X atténués par le corps humain pour les retranscrire sur un film ou de manière digitale. Des filtres sont souvent ajoutés à la sortie de la source pour que les rayons X de faible énergie viennent s'y collimater, ce qui améliore le rapport signal à bruit final.

2.2.2 Système EOS

Le système EOS (EOS imaging, Paris, France), illustré à la Figure 2.6, est une méthode récente d'acquisition d'images médicales à partir de rayons X qui permet d'obtenir des images biplans dans les plans coronal et sagittal avec une possibilité de reconstruction en trois dimensions de certaines structures osseuses (Dubousset et al., 2005). Il est aujourd'hui considéré comme le nouvel étalon-or pour l'évaluation des structures ostéo-articulaires (Illés and Somoskeöy, 2012; Wybier and Bossard, 2013).

Le principe physique du système d'acquisition d'images EOS repose sur les chambres à fils, ou chambres proportionnelles multifilaires, proposées par Georges Charpak, invention qui lui

(a) Cage du système EOS. (b) Intérieur du système EOS.

Figure 2.6 Système EOS, images issues de EOS Espace média

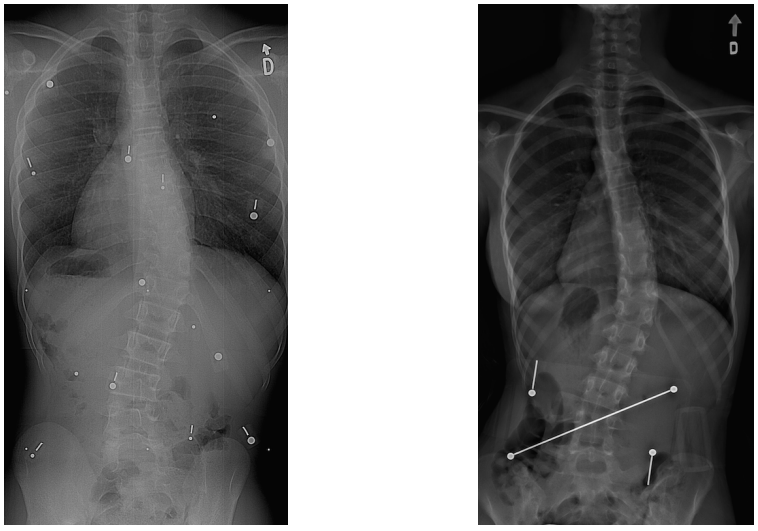
a valu le Prix Nobel de physique en 1992. Une chambre à fils contient un gaz noble sous pression tel que le xénon pour convertir les photons X en électrons (Illés and Somoskeöy, 2012). Lorsqu'un photon entre dans la chambre à fils, il percute les atomes neutres du gaz présent –on parle d'ionisation– ce qui crée alors des électrons et des cations. Le nombre d'électrons libérés dépend de l'intensité du photon. Ces électrons libres sont alors accélérés dans un champ électrique. Un effet d'avalanche survient lorsque les électrons libres acquièrent un niveau d'énergie suffisant, grâce au champ électrique, pour percuter de nouveaux atomes neutres du gaz. Ce faisceau d'électrons créé par effet d'avalanche va alors générer un courant électrique sur les grilles de fils présentes dans la chambre. La variation du potentiel des fils permet alors de compter le nombre de photons, quasiment au photon près, avec une grande sensibilité, et de localiser leur trajectoire.

L'instrumentation du système EOS ressemble à un système de radiographie classique à la différence qu'une chambre à fils se situe entre le patient et le détecteur de rayon X. À cela s'ajoute également le mécanisme d'acquisition qui consiste en une cage en forme de L, illustré à la Figure 2.6. Ces particularités propres au système EOS engendrent de nombreux avantages par rapport à la radiologie classique :

— La diffusion des rayons X dans l'air est réduite ce qui permet une meilleure collimation des faisceaux à rayons X sur la personne imagée (Wybier and Bossard, 2013).
— Les erreurs de parallaxe sont évités puisque le système est capable d'émettre et de détecter les rayons X sur les plans coronal et sagittal en même temps (Wybier and Bossard, 2013).
— L'exposition aux rayons ionisants est 8 à 10 fois moindre par rapport à la radiologie bidimensionnelle et de 800 à 1000 fois moindre par rapport à la tomodensitométrie

tridimensionnelle (Dubousset et al., 2005).

Ces propriétés impactent directement la qualité des images produites qui possèdent notamment un rapport signal à bruit supérieur à la radiologie classique (Wybier and Bossard, 2013), comme illustré à la Figure 2.7. De plus, la résolution est de l'ordre de 250 µm et un grand nombre de niveaux de gris (de 30 à 50 000) sont discernables (Dubousset et al., 2005). Ces caractéristiques font en sorte que le système EOS devient de plus en plus attractif par rapport à la radiologie classique puisqu'il produit des images de plus grande qualité avec une exposition aux rayons ionisants inférieure. L'inconvénient principal provient du temps d'acquisition de l'ordre de 30 à 45 sec, qui fait en sorte que plusieurs prises sont parfois nécessaires suite au mouvement du patient (Wybier and Bossard, 2013).



(a) Radiographie classique. (b) Radiographie système EOS.

Figure 2.7 Comparaison qualitative entre deux radiographies (par projection et système EOS)

L'autre avantage majeur, qui a contribué au succès du système EOS, concerne la partie logicielle qui a été développée conjointement entre le Laboratoire de biomécanique (LBM) de l'École Nationale Supérieure des Arts et Métiers (Paris, France) et le Laboratoire de recherche en imagerie et orthopédique (LIO) de l'École de Technologie Supérieure (Montréal, Canada) (Dubousset et al., 2005). Les méthodes développées permettent notamment de reconstruire précisément la colonne vertébrale en trois dimensions à partir des radiographies biplanaires du système EOS (Pomero et al., 2004). D'autres structures osseuses ont par la suite été prises en compte comme le bassin, le genou, le fémur, ou encore le tibia (Illés and Somoskeöy, 2012). Le principe consiste à construire un modèle statistique de la structure osseuse d'intérêt à partir d'une large base de données. Lorsqu'un nouveau patient est imagé,

l'algorithme de reconstruction repère dans un premier temps les contours des structures osseuses sur les plans coronal et sagittal. Le modèle statistique est par la suite déformé par des transformations affines puis projetés sur les plans coronal et sagittal. Il s'agit d'un processus itératif qui va chercher à minimiser l'erreur des contours réels (c'est-à-dire ceux qui sont imagés) et virtuels (c'est-à-dire ceux qui proviennent du modèle statistique).

2.2.3 Tomodensitométrie

La tomodensitométrie est une méthode d'acquisition imagerie par rayons X qui permet d'obtenir des coupes virtuelles au lieu des projections habituellement obtenues par la radiographie classique ou par le système EOS. Le principe général consiste en une source de rayons X qui tourne autour du patient. Une étape supplémentaire qui implique un traitement numérique est requise pour reconstruire une image à partir des signaux détectés.

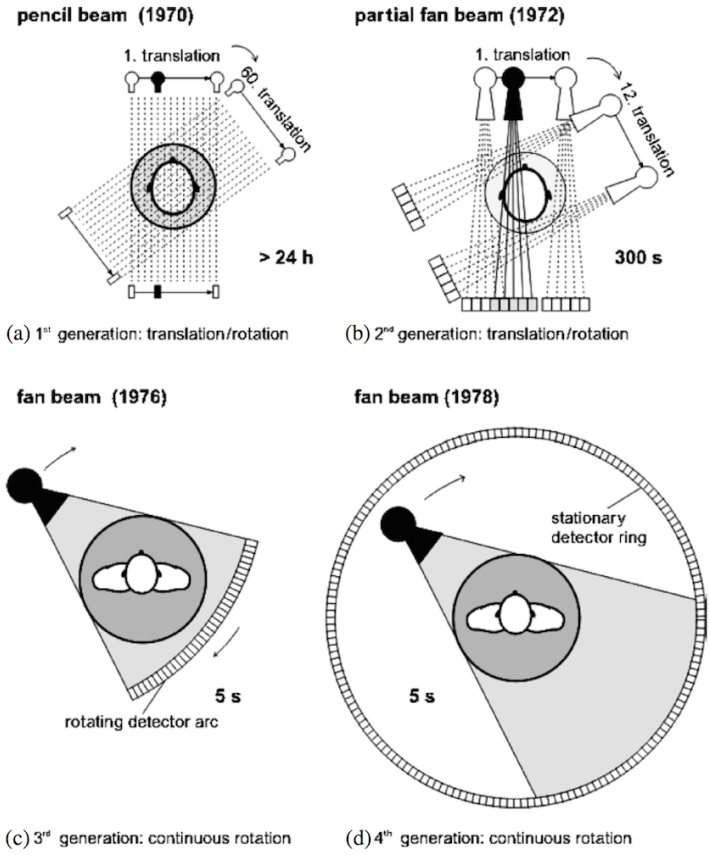


Figure 2.8 Illustration des quatre générations de tomodensitométrie, tirée de Kalender (2006).

Les premiers travaux sur la tomodensitométrie date de 1972 par Godfrey N. Hounsfield qui était alors ingénieur à la maison de disques EMI Ltd, invention qui lui a valu le Prix Nobel

de physique en 1978. Les profits générés par les Beatles ont en effet financé les recherches de Hounsfield (Goodman, 2010). Au fil des années, la méthode d'acquisition a constamment évolué pour réduire le temps d'acquisition et l'exposition aux rayons ionisants. La Figure 2.8 illustre les quatre générations de tomodensitométrie proposée :

— La première génération correspond à celle proposée initialement par Hounsfield. Le corps à imager est irradié par une source qui émet un fin faisceau de rayons X. Un détecteur de l'autre côté du corps détecte les rayons X transmis. Une première mesure est alors obtenue. Ce processus est ensuite répété pour différents pas afin couvrir tout le corps par translation. À ce stade-ci, les données acquises ne permettent que d'avoir une image de projection. Il faut alors ajouter des rotations du système source-détecteur pour au moins obtenir 180 projections afin de reconstruire une coupe (Fanet, 2010).
— La seconde génération reprend le concept de translation-rotation et correspond aux premiers systèmes de tomodensitométrie commercialisés (Fanet, 2010). Au lieu d'avoir un détecteur unique, une barrette de détecteurs est utilisée en conjonction avec un faisceau en éventail. Ce système a alors permis de réduire le temps d'acquisition d'une coupe de plus de 24 h à 300 sec (Kalender, 2006).
— La troisième génération a étendu la notion de faisceau en éventail par l'utilisation d'un grand angle d'ouverture. Une grande section pouvant être irradiée, la translation n'est alors plus nécessaire ce qui a encore réduit drastiquement le temps d'acquisition pour obtenir une coupe (Fanet, 2010).
— La quatrième génération comporte un anneau complet de détecteurs entourant le patient. Seule la source de rayons X est en rotation. Toutefois, ce système demeure très sensible aux artéfacts liés aux effets de diffusion des rayons X (Fanet, 2010).

La tendance aujourd'hui s'oriente vers une amélioration de la troisième génération avec une source émettant un faisceau conique et des détecteurs matriciels pour imager plusieurs coupes en même temps (Grangeat, 2002; Fanet, 2010). De plus, le temps d'acquisition a été réduit en faisant bouger le lit du patient en même temps que la rotation du système source-détecteur, créant une acquisition hélicoïdale capable d'imager le patient en 30 sec (Grangeat, 2002).

Cependant, l'ensemble des systèmes physiques de la tomodensitométrie ne permet que d'acquérir les projections des rayons X. Une reconstruction numérique est par la suite nécessaire pour construire la cartographie d'absorption des rayons X. Il s'agit en effet d'un problème d'inversion qui peut être résolu par des méthodes analytiques ou par des méthodes algébriques, qui sortent du cadre de ce mémoire.

L'image obtenue par tomodensitométrie représente ainsi une cartographie de l'absorption des rayons X par le corps humain. Cette absorption suit la loi de Beer-Lambert (Éq. 2.1)

mais dans sa forme généralisée à des matériaux inhomogènes (Fanet, 2010). Au final, chaque pixel, ou plutôt voxel, de l'image s'exprime en unité de Hounsfield (HU) qui correspond au coefficient d'atténuation linéaire normalisé par les coefficients de l'eau ($\mu_{eau} = 0$) et de l'air ($\mu_{air} = -1000$) :

$$HU = 1000 \times \frac{\mu - \mu_{eau}}{\mu_{eau} - \mu_{air}} \tag{2.2}$$

Comme expliqué à la section 2.2.1, la valeur de chacun des voxels dépend du numéro atomique du matériau imagé. Chaque voxel est donc associé à une représentation fidèle des propriétés physiques du matériau. La tomodensitométrie génère donc des images qui ont une gamme de niveaux de gris très large pour couvrir les différents coefficients d'absorption observés. Les radiologues utilisent souvent un fenêtrage des niveaux de gris pour ajuster leur œil à la région d'intérêt. Ce principe de fenêtrage sera d'ailleurs exploité au cours des travaux portant sur la segmentation de reins au chapitre 6.

La tomodensitométrie est également pertinente pour l'étude de la vascularisation des organes grâce à l'utilisation d'agent de contraste à base d'iode. L'agent de contraste injecté en intra-veineuse au sujet se propage dans le corps humain. La tomodensitométrie acquiert des images en concordance avec cette propagation pour obtenir un meilleur contraste dans les tissus mous. Comparé à l'imagerie par résonance magnétique, la tomodensitométrie avec agent de contraste offre une résolution supérieure pour un temps d'acquisition moindre. C'est pourquoi la tomodensitométrie reste encore populaire, malgré les rayons ionisants, pour des applications cérébrales, thoraciques ou encore abdominales (Grangeat, 2002) qui sont de notre intérêt pour le chapitre 6.

2.3 Intérêts de l'apprentissage automatique

Lorsque les jeux de données sont relativement petits, des algorithmes peuvent être construits "manuellement" par des règles d'association. Toutefois, lorsque le jeu de données devient imposant, de nouveaux défis s'imposent. Il devient en effet plus ardu de capturer la variabilité au sein du jeu de données par de simples règles d'association. L'apprentissage automatique vient alors pallier ce problème en apprenant des tendances présentes dans le jeu de données. Plus particulièrement, l'apprentissage de représentations, basé sur des réseaux de neurones artificiels, cherche à démêler les facteurs de variation au sein du jeu de données pour apprendre des traits caractéristiques discriminants pour la tâche que l'on souhaite accomplir. Dans le cadre de ce mémoire, la tâche visée concerne la classification d'images biomédicales. Concrètement, cela se traduit dans un premier temps, par la découverte de patrons de pathologies au sein

une large base de données multicentriques de patients atteints de la scoliose idiopathique de l'adolescent (AIS) ; et dans un second temps, par la découverte de traits caractéristiques discriminants pour la segmentation de reins au sein d'une large base de données de patients atteints d'anévrisme de l'aorte abdominale.

2.3.1 Classification de la scoliose idiopathique de l'adolescent

Le cas des systèmes de classification de l'AIS illustre ce besoin d'aller au-delà de simples règles d'association. Les systèmes actuellement utilisés par les chirurgiens orthopédiques se basent sur des mesures en deux dimensions (2D) extraites de radiographies dans les plans coronal et sagittal (King et al., 1983; Lenke et al., 2001), qui proviennent de systèmes classiques (section 2.2.1) ou de systèmes EOS (section 2.2.2). Or, des profils de courbure de la colonne vertébrale peuvent apparaître similaires sur les plans coronal et sagittal mais différer complètement lorsque que l'on considère la déformation en 3D (Labelle et al., 2011).

Dans la littérature, de nombreux indices géométriques ont été proposés pour décrire quantitativement la scoliose en 3D pour améliorer les traditionnelles métriques comme l'angle de Cobb (décrit à la section 2.1.1). Poncet et al. (2001) ont introduit une métrique basée sur la torsion géométrique de la colonne vertébrale. Kadoury et al. (2014) ont étendu ces travaux pour caractériser la torsion au niveau des courbures régionales, ce qui a permis de découvrir des nouveaux sous-groupes par un algorithme de partitionnement en k-moyennes floues. Sangole et al. (2009) ont inclus des mesures régionales (décrits à la section 2.1.1) et utilisé l'algorithme ISOData (une variante des k-moyennes) pour subdiviser des courbures thoraciques classifiées de type Lenke-1. Duong et al. (2009) ont combiné les indices de torsion et ceux issus du plan de courbure maximale pour produire des sous-groupes par l'algorithme des k-moyennes. Globalement, ces articles partagent un cadre similaire au sein duquel des indices sont extraits à partir de reconstructions de colonne vertébrale pour partitionner la cohorte en différents sous-groupes cliniquement significatifs. Toutefois, se baser essentiellement sur des indices géométriques pose le problème d'extraction de traits caractéristiques, c'est-à-dire la manière dont la reconstruction de colonne vertébrale doit être représentée.

Les algorithmes devraient au contraire pouvoir capturer la dimension intrinsèque de ces reconstructions de colonne vertébrale pour modéliser les changements globaux (au niveau du rachis) et locaux (au niveau des vertèbres). Duong et al. (2006) ont proposé une décomposition en ondelettes de la reconstruction de la colonne vertébrale. Kadoury and Labelle (2012) ont investigué l'utilisation d'un algorithme pour la réduction de dimensionnalité par *linear local embedding*. Cependant, ces méthodes locales tendent à souffrir du fléau de la dimensionnalité ce qui les rend sensibles aux cas aberrants (van der Maaten et al., 2009). Par

conséquent, si les reconstructions de la colonne vertébrale sont plus précises par l'ajout de marqueurs supplémentaires ou si les patients au sein du jeu de données présentent de trop fortes variations, alors ces méthodes produiront des mauvaises classifications.

Une méthode globale et non-linéaire d'apprentissage non-supervisée basée sur des auto-encodeurs empilés est proposée au cours du chapitre 5 pour subvenir à ces problèmes en préservant les propriétés globales des reconstructions de la moelle épinière.

2.3.2 Classification de voxels d'images tomodensitométriques pour la segmentation des reins

La segmentation d'organes fait également face au besoin de développer des algorithmes qui puissent se généraliser à de nouveaux cas. Cela est d'autant plus important que les images obtenues par tomodensitométrie comprennent de nombreux artefacts liés au bruit électronique de l'appareil d'acquisition, de la très forte variabilité intra- et inter-sujets, mais aussi des différences intrinsèques à chaque manufacturier d'appareil de tomographie.

À travers la littérature, les approches par recalage d'images pour la segmentation d'organes sont monnaie courante en imagerie médicale, particulièrement pour les structures cérébrales. L'idée est de recaler un atlas de l'organe d'intérêt sur l'image à segmenter. Cette procédure intuitive repose fortement sur la qualité de l'atlas. Les performances de généralisation demeurent le plus souvent faibles (Criminisi et al., 2013) à cause de la forte variabilité entre les patients dont les pathologies ne se manifestent pas de la même manière. Pour compenser ce défaut, Wolz et al. (2012) ont proposé une technique basée sur de multiples atlas à différentes échelles. Cette approche multiéchelle produit la segmentation d'organes de la région abdominale, en pondérant l'importance de chaque atlas par l'apparence de l'image et de l'organe en tant que tel. Chu et al. (2013) ont étendu cette approche en divisant l'image en de multiples sous-espaces. Un graphe est construit pour chacun de ces espaces. La segmentation est alors obtenue en coupant le graphe en deux sous-ensembles. Cependant, le principal défaut des approches par recalage d'images multiéchelles provient des nombreux recalages qui demandent plusieurs heures de calcul pour produire les segmentations (Wolz et al., 2012; Chu et al., 2013).

Des approches par des champs aléatoires ont également été proposées car les modèles incorporent de l'information liée à la structure des voxels, c'est-à-dire la dépendance d'un voxel avec ses voisins. Freiman et al. (2010) ont utilisé une approche par coupe minimum pour segmenter automatiquement les reins. Khalifa et al. (2011) ont proposé une approche par ligne de niveau avec un a priori provenant de champs aléatoires. Cependant, les approches par des champs aléatoires sont des méthodes itératives qui nécessitent d'échantillonner un

graphe de grande dimension ce qui prend également des temps de calculs importants.

Avec une disponibilité grandissante des images médicales, des approches par apprentissage automatique basées sur des forêts aléatoires sont devenues attractives pour la détection (Criminisi et al., 2013) et la segmentation d'organes. Cuingnet et al. (2012) ont proposé une approche avec des forêts aléatoires pour régresser les coordonnées des boîtes englobant le rein, suivie de forêts aléatoires pour classifier chacun des voxels de la boîte englobante avant de déformer un modèle en forme d'ellipsoïde pour obtenir la segmentation des reins. Glocker et al. (2012) ont étendu cette idée en combinant la régression et la classification pour la segmentation de multiples organes dans l'abdomen par une nouvelle fonction objectif qui permet aux forêts aléatoires d'apprendre la classe du voxel et sa position spatiale. Cependant, ces approches par forêts aléatoires dépendent des traits caractéristiques spécifiés par l'utilisateur.

Dans le cadre de ce mémoire, une méthode d'apprentissage supervisée basée sur des réseaux à convolution sera proposée au cours du chapitre 6 pour segmenter les reins dans un jeu de données à partir de traits caractéristiques appris durant l'entrainement.

Au vu des points énoncés tout au long de cette section, les sections suivantes de la revue de littérature présentent l'état de l'art concernant la construction et l'entrainement d'un réseau de neurones. Les méthodes exposées sont à la base des méthodologies utilisées dans les travaux des chapitres 4, 5 et 6.

2.4 Modèles linéaires et objectifs d'apprentissage

Une fonction objectif $L(x, y)$, également appelée fonction de perte, doit être définie pour mesurer l'écart entre la prédiction basée sur l'entrée x et la véritable valeur y. Les paramètres de l'algorithme sont par la suite modifiés dans le but de minimiser l'erreur définie par cette fonction objectif. L'ensemble des erreurs de l'algorithme se traduit par l'évaluation d'un risque :

$$R(f) = E[L(x, y)] = \int L(x, y) dP(x, y) \tag{2.3}$$

où f est la fonction apprise par l'algorithme, $E[\cdot]$ l'espérance et $P(x, y)$ la loi de probabilité à plusieurs variables décrivant les données. L'objectif ultime de l'apprentissage est alors de trouver une fonction f qui minimise le risque $R(f)$. Or, $P(x, y)$ est inconnue. Une approximation peut toutefois en être déduite à partir du jeu de données d'entrainement $\mathcal{D}_{train}$ qui est à notre disposition. L'hypothèse principale repose sur le fait que les observations de $\mathcal{D}_{train}$, et celles des autres partitions, sont des variables indépendantes et identiquement distribuées

qui proviennent de $P(x, y)$. On parle alors de risque empirique :

$$\hat{R}(f, \mathcal{D}_{train}) = \frac{1}{N} \sum_{i=1}^{N} L(x^{(i)}, y^{(i)}) \tag{2.4}$$

où N représente le nombre d'exemples présents dans le jeu de données. La fonction apprise par l'algorithme minimise le risque empirique et se dénote par $\hat{f}_N = \arg\min_{f \in F} \hat{R}(f, \mathcal{D}_{train})$, telle que mentionnée à la section 1.1.2.

Prenons maintenant deux méthodes linéaires pour la régression et la classification, que sont respectivement la régression linéaire et la régression logistique.

2.4.1 Régression linéaire

Dans le cadre d'une régression linéaire, l'objectif est de trouver les paramètres W et b, appelés respectivement *poids* et *biais*, de la fonction f pour prédire la variable $y \in \mathbb{R}$ dépendante des données x. Cette prédiction s'exprime par :

$$\hat{y} = Wx + b \tag{2.5}$$

Lors de l'apprentissage, la fonction objectif mesure l'erreur quadratique :

$$L(x, y) = \|\hat{y} - y\|^2 \tag{2.6}$$

Ainsi, le risque empirique correspond à l'erreur quadratique moyenne :

$$\hat{R}(f, \mathcal{D}_{train}) = \frac{1}{N} \sum_{i=1}^{N} \|\hat{y}^{(i)} - y^{(i)}\|^2 \tag{2.7}$$

2.4.2 Régression logistique

La régression logistique est une extension de la régression linéaire pour des catégories. L'objectif est de trouver les paramètres W et b de la fonction f pour prédire la variable $y \in \{0, 1\}$ dans le cas binaire, ou $y \in \{1, \cdots, L\}$ dans le cas multiclasse, dépendante des données x.

Classification binaire

Contrairement à l'Éq. 2.5, une non-linéarité est ajoutée pour obtenir une probabilité entre 0 et 1 d'appartenir à la classe 1. Cette non-linéarité s'appelle la fonction logistique :

$$P(y = 1|x, \theta) = \hat{y} = \frac{1}{1 + \exp\left(-(W^T x + b)\right)} \tag{2.8}$$

où $\theta = \{W, b\}$. Lors de l'apprentissage, la fonction objectif mesure l'entropie croisée binaire :

$$L(x, y) = y \log \hat{y} + (1 - y) \log\left(1 - \hat{y}\right) \tag{2.9}$$

La moyenne des entropies croisées sur l'ensemble d'entrainement produit l'estimateur du risque empirique :

$$\hat{R}(f, \mathcal{D}_{train}) = \frac{1}{N} \sum_{i=1}^{N} y^{(i)} \log\left(P(y^{(i)} = 1|x^{(i)}, \theta)\right) + (1 - y^{(i)}) \log\left(1 - P(y^{(i)} = 1|x^{(i)}, \theta)\right) \tag{2.10}$$

Classification multiclasses

Lorsque le nombre de classes est supérieur à 2, la fonction softmax est nécessaire pour estimer la probabilité de chacune des classes. Il est à noter que la fonction softmax est une généralisation de la fonction logistique à plusieurs classes.

$$P(y = i|x, \theta_i) = \hat{y}_i = \frac{\exp\left(W_i^T x + b_i\right)}{\sum_{j=1}^{L} \exp\left(W_j^T x + b_j\right)} \tag{2.11}$$

où $\theta_i = \{W_i, b_i\}$. Lors de l'apprentissage, la fonction objectif mesure l'entropie croisée pour plusieurs catégories :

$$L(x, y) = -\log \hat{y}_i \tag{2.12}$$

La moyenne des entropies croisées sur l'ensemble d'entrainement produit l'estimateur du risque empirique :

$$\hat{R}(f, \mathcal{D}_{train}) = \frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)}|x^{(i)}, \theta) \tag{2.13}$$

Cependant, cette forme de la fonction softmax peut se révéler instable lorsque les nombres à virgules sont représentés en simple précision, ce qui est notamment le cas lors des opérations effectuées par un processeur graphique. Un dépassement en virgule flottante peut en effet survenir à cause de la fonction exponentielle. La solution est de retrancher la valeur maximale de l'argument lors de l'implémentation de la fonction softmax (Bengio et al., 2015). De plus,

un soupassement en virgule flottante peut également survenir lorsque le risque empirique est calculé, c'est-à-dire lorsque l'on prend le logarithme de la probabilité $\hat{y}_i$.

2.5 Réseau de neurones artificiels

Les réseaux de neurones artificiels (ANN) tirent leur inspiration des neurones biologiques. Leur architecture reprend notamment des concepts similaires. Toutefois, leur mode de fonctionnement est diamétralement opposé et leur performance demeure, encore à ce jour, loin des neurones biologiques. Il est à noter que l'apprentissage de représentations s'inspire de la biologie mais n'a pas pour objectif de répliquer ou de comprendre le fonctionnement des neurones biologiques. Ce domaine d'études concerne les neurosciences computationnelles.

Cette section décrit dans un premier temps le fonctionnement d'un neurone biologique standard avant d'aborder les ressemblances avec un ANN. Plusieurs architectures de ANN utilisées dans les travaux de ce mémoire sont par la suite détaillées.

2.5.1 Analogie avec la neurobiologie

Le neurone constitue l'unité fonctionnelle de base du système nerveux dont la principale fonction concerne la transmission de l'information (Hall, 1994). À travers ses récepteurs, le neurone reçoit de l'information qu'il traite pour envoyer à son tour un message aux autres neurones, ou à d'autres tissus biologiques comme les muscles. Le nombre de neurones dans le système nerveux humain est estimé à plus de 100 milliards, et chaque neurone est capable de traiter l'information provenant de plus de 10 000 neurones différents (Godaux, 1994). Cette interaction complexe entre les neurones est à l'origine des processus cognitifs complexes de l'être humain, tels que la mémoire, la perception sensorielle, le langage, ou encore la planification motrice et spatiale. Pour ce faire, le neurone possède différents aspects – taille du corps cellulaire, forme, organisation cellulaire – résultant en plus de 10 000 classes différentes pour répondre à des fonctions très spécifiques (Hall, 1994). Des topologies particulières existent par exemple pour les systèmes sensoriels. Dans le cadre de ce mémoire, nous nous attèlerons seulement à présenter le neurone dit "standard" pour illustrer l'analogie entre les neurones biologique et artificiel. La Figure 2.9 illustre les principales parties d'un neurone biologique.

L'architecture d'un neurone présente une morphologie particulière dont chacune des composantes remplit un rôle particulier. Le neurone est formé d'un corps cellulaire comprenant un noyau cellulaire. À partir de ce corps cellulaire, parfois appelé soma, de nombreuses extensions en forme de branches d'arbre, appelées dendrites, se déploient pour recevoir l'information d'autres neurones grâce à leur récepteur synaptique. Le soma traite alors les signaux

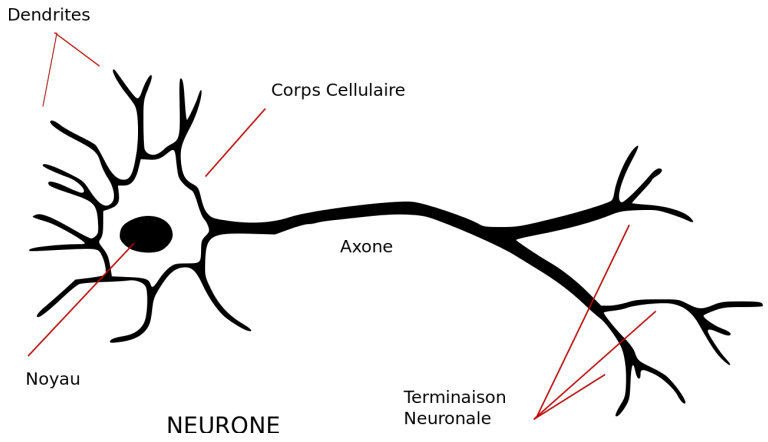


Figure 2.9 Illustration d'un neurone biologique, tirée de Wikimedia Commons (2007).

reçus au niveau des récepteurs des dendrites pour les combiner ou traduire en d'autres signaux électriques, créant un potentiel synaptique global. Une autre extension plus longue et sans subdivisions, appelée axone, se charge de transporter cette nouvelle information du soma vers les autres neurones. À l'extrémité de l'axone, de nombreuses ramifications se déploient pour orienter cette nouvelle information vers les dendrites des neurones subséquents. De plus, l'axone se caractérise par son fonctionnement en impulsion électrique qui lui permet de transmettre l'information rapidement aidé par la gaine de myéline qui l'entoure. La jonction entre la terminaison neuronale et le neurone recevant l'information s'appelle une synapse. C'est à cet endroit que la transmission concrète d'information se fait par l'intermédiaire de neurotransmetteurs. Le neurone post-synaptique, c'est-à-dire celui qui reçoit les neurotransmetteurs, peut alors recevoir un signal d'excitation ou d'inhibition selon les composants relargués par la synapse. La conduction de l'information se fait par des potentiels électriques. Le neurone possède une différence de potentiel membranaire négative au repos qui se dépolarise lorsque qu'un potentiel d'action est déclenché. Pour qu'une activation survienne, il faut que les stimuli dépassent un certain seuil. Le potentiel d'action, qui possède toujours la même amplitude quelle que soit l'intensité du stimulus, se propage alors tout au long de l'axone jusqu'au relargage des neurotransmetteurs. Une période de répit, appelée période réfractaire, empêche d'envoyer une succession de potentiels d'action pour assurer le contrôle de l'information. Un neurone consiste donc une unité du système nerveux capable de recevoir en entrée de l'information par les récepteurs présents sur ses dendrites, traiter les signaux reçus dans le soma et finalement envoyer un signal aux neurones subséquents de manière précise, rapide et coordonnée.

2.5.2 Réseau de neurones artificiels

Un réseau de neurones artificiels (ANN) est composé de plusieurs couches, tout comme des neurones biologiques qui se suivent en série ou en parallèle. Chacune des couches possède des paramètres θ pour transformer les données reçues en un signal de sortie pour la couche suivante. Ces paramètres peuvent être soit négatifs, soit positifs. En d'autres mots, cette configuration ressemble au neurone biologique qui reçoit des signaux par ses dendrites, effectue une combinaison au niveau du soma pour produire un signal d'inhibition ou d'excitation qui est envoyé au neurone subséquent par son axone. De plus, une non-linéarité, s'inspirant du seuil de déclenchement du neurone biologique, peut être ajoutée à la transformation affine des entrées par les paramètres.

Plus formellement, un ANN comprend plusieurs couches L. Chaque couche est désignée par $l \in \{0, \cdots, L-1\}$ où la couche 0 correspond au vecteur d'entrée $x = y^{(0)}$ et la couche $L-1$ à la cible de sortie $y = y^{(L-1)}$ à prédire. Les couches entre celle d'entrée et celle de sortie sont dites *cachées* ou *latentes* car il s'agit des couches pour lesquelles que l'ANN apprend les paramètres. Selon la tâche qui souhaite être apprise, la couche de sortie correspond à une régression linéaire (section 2.4.1) ou à une régression logistique (section 2.4.2). Le mécanisme d'apprentissage d'un ANN consiste à estimer les poids $W^{(l)}$ et les biais $b^{(l)}$ de chaque couche cachée l. La transformation du vecteur d'entrée x en la cible de sortie y consiste alors en une succession d'opérations :

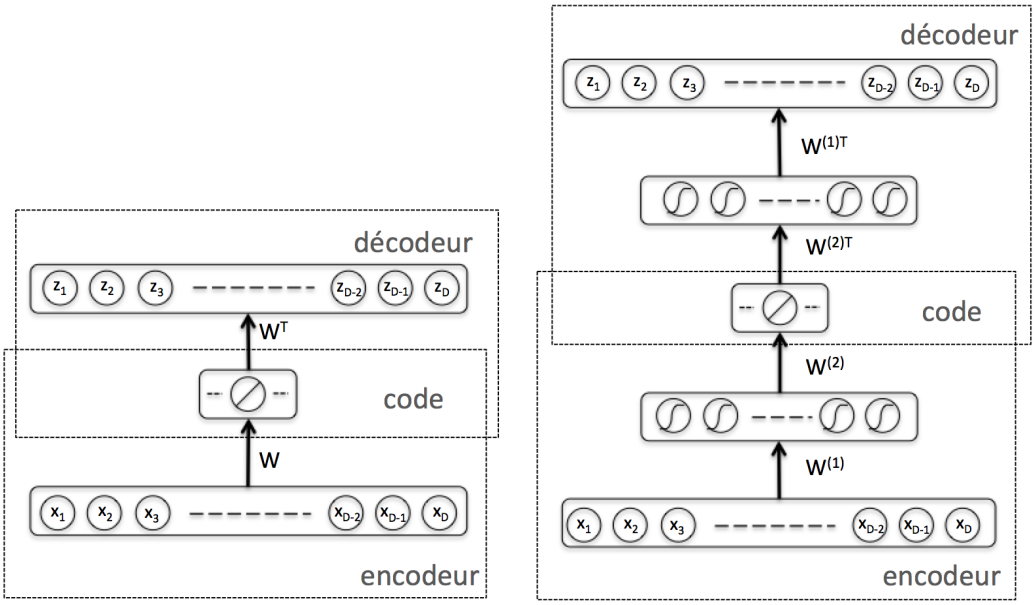
$$y^{(l)} = s(W^{(l)} y^{(l-1)} + b^{(l)}) \tag{2.14}$$

où $s(\cdot)$ correspond à la fonction d'activation. Les non-linéarités les plus courantes utilisées pour la fonction d'activation sont décrites à la section 2.6.2.

2.5.3 Auto-encodeur

Un auto-encodeur peut simplement être défini comme un réseau de neurones artificiels qui apprend une représentation cachée dans le but de reconstruire ses entrées. Il s'agit donc d'une architecture non-supervisée où les entrées du modèle sont similaires aux sorties, et où le modèle optimise ses paramètres selon l'erreur quadratique moyenne (voir Éq. 2.7). Un auto-encodeur peut-être utilisé à des fins de pré-entrainement (décrit à la section 2.6.1) ou à des fins de réduction de dimensionnalité lorsque l'architecture prend la forme d'un goulot d'étranglement.

À des fins de simplicité, considérons un auto-encodeur à une couche cachée ; l'auto-encodeur pourra par la suite apprendre plusieurs niveaux de représentations en empilant des couches cachées. Dans un premier temps le vecteur d'entrée x de dimension D est transformé par



(a) Auto-encodeur. (b) Auto-encodeurs empilés.

Figure 2.10 Auto-encodeurs à une (a) ou plusieurs (b) couches cachées.

un fonction d'encodage f vers la couche cachée h (souvent appelée *code* dans le cas des auto-encodeurs) :

$$h = f(x) = s(W^{(1)}x + b^{(1)}) \tag{2.15}$$

où $W^{(1)}$ est la matrice des paramètres de la fonction d'encodage, $b^{(1)}$ le vecteur de biais et $s(\cdot)$ la fonction d'activation. Il est à noter qu'un auto-encodeur à une couche cachée correspond à une analyse en composantes principales si la fonction d'activation est linéaire. Une fois le code produit, une fonction de décodage g est appliquée pour revenir à la dimension du vecteur d'entrée x ce qui résulte en un vecteur de reconstruction z :

$$z = g(f(x)) = s(W^{(2)}h + b^{(2)}) \tag{2.16}$$

où $W^{(2)}$ est la matrice des paramètres de la fonction de décodage. Contraindre la transposée de $W^{(1)}$ à être égale à $W^{(2)}$ –c'est-à-dire $W^{(2)} = W^{(1)^T}$– offre de nombreux avantages. Cela joue un rôle de régulariseur : en évitant des mises à jour des poids par de faibles valeurs, et en réduisant le nombre de paramètres à optimiser (Bengio et al., 2013). Dans le cas de la réduction de dimensionnalité, le code h possède nécessairement une plus faible dimensionnalité que le vecteur d'entrée x.

2.5.4 Réseau à convolution

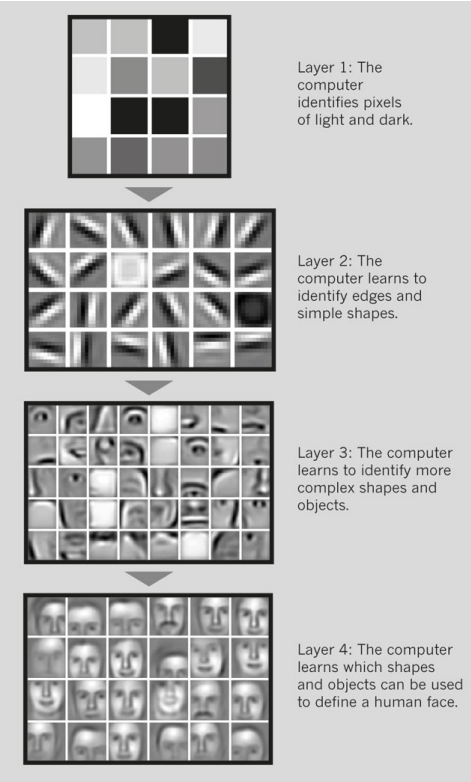


Figure 2.11 Reconnaissance de visages avec un ConvNet, image tirée de Jones (2014).

Les réseaux à convolution (ConvNets) ont montré des performances de généralisation remarquables sur des jeux de données très larges comprenant plusieurs millions d'images (Russakovsky et al., 2014; Krizhevsky et al., 2012; Sermanet et al., 2013; Simonyan and Zisserman, 2014). Ces succès proviennent principalement de l'architecture particulière des ConvNets qui tient compte de la topologie spécifique des tâches reliées à la vision par ordinateur qui impliquent des images en deux dimensions. D'autres dimensions peuvent également être prises en compte lorsqu'il s'agit par exemple d'images en couleurs avec plusieurs canaux.

Les ConvNets exploitent la très forte corrélation qu'il existe au sein d'une structure locale en deux dimensions en restreignant le champ récepteur des unités cachées à se focaliser sur des variations locales (LeCun et al., 1998). Ainsi, un schéma de connectivité locale est appris au cours des premières couches cachées pour décrire des structures simples telles que des bords ou des coins. Empiler des couches de convolution, c'est-à-dire aller en profondeur, force le réseau à apprendre des représentations plus abstraites et plus discriminantes en combinant au sein des couches plus profondes les traits caractéristiques locaux appris au cours des premières couches (LeCun et al., 1998).

Prenons un exemple concret pour illustrer les représentations apprises par un ConvNet. Dans cet exemple, la tâche d'apprentissage est la reconnaissance de visages qui est représentée à la Figure 2.11. Le ConvNet doit donc apprendre à discriminer les visages des autres objets de l'image. Au sein de la première couche cachée (*Layer 2* dans la Figure 2.11), les filtres détectent des formes simples comme les bordures, et ce selon plusieurs rotations. Ces filtres sont très similaires aux filtres de Gabor utilisés en vision par ordinateur. Plus la couche cachée est profonde, plus les traits caractéristiques appris tendent à être complexes et abstraits. À la seconde couche cachée, les filtres résultent d'une combinaison des filtres précédents pour former motifs plus complexes qui correspondent à des parties spécifiques du visage, comme les yeux, le nez ou la bouche. Enfin, la dernière couche cachée montre plusieurs représentations abstraites d'un visage humain.

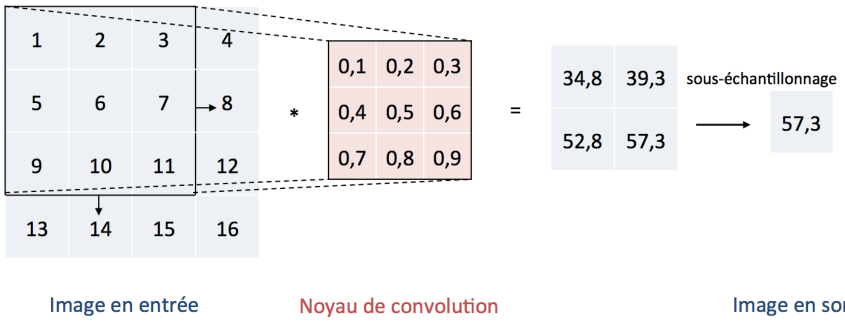


Figure 2.12 Principe d'une couche de convolution. Un noyau de convolution de taille 3×3 est appliqué à toute l'image de taille 4×4 avec des poids similaires par une technique de fenêtre glissante. Une image de taille 2×2 est alors produite. Un sous-échantillonnage de taille 2×2 où seule la valeur maximale est conservée génère l'image de sortie qui dans cet exemple est de taille 1×1.

De plus, les ConvNets sont invariants à la translation à cause de plusieurs contraintes spécifiques à leur architecture. Les unités d'une couche de convolution sont organisées dans une topologie en deux dimensions, appelée un noyau de convolution. Une couche de convolution peut également comporter plusieurs noyaux de convolution. Le terme de *filtre* est également souvent utilisé pour désigner un noyau de convolution. Les paramètres d'un noyau sont partagés, ce qui veut dire que la même opération est appliquée sur l'image d'entrée par une technique de fenêtre glissante. Le fait de partager les poids pousse le réseau à devenir invariant à la translation puisque les traits caractéristiques appris dans le noyau seront appliqués à toute l'image et non à une portion spécifique de l'image. À cela s'ajoute une couche de sous-échantillonnage, qui renforce d'autant plus cette propriété. Une couche de sous-échantillonnage effectue une opération de décimation locale où seule la valeur maximale de l'image résultant de l'opération de convolution est conservée. Ces trois principes

(connectivité locale, partage des paramètres, sous-échantillonnage) forment le cœur d'une couche de convolution dans un ConvNet. Ces opérations sont représentées visuellement à la Figure 2.12. De nos jours, une couche de convolution contient typiquement une (ou plusieurs) opération de convolution suivie d'un sous-échantillonnage. L'empilement de plusieurs couches de convolution crée un ConvNet.

2.6 Entraînement d'un réseau de neurones artificiels

2.6.1 Optimisation

Descente de gradient stochastique

La descente de gradient stochastique est une méthode d'optimisation couramment utilisée aujourd'hui pour entrainer des ANNs. Il s'agit d'un processus itératif au sein duquel l'algorithme d'apprentissage ajuste tout au long de l'entrainement ses paramètres. À chaque mise à jour, des exemples sont présentés au ANN qui va faire une prédiction et estimer l'erreur associée. Un ou plusieurs exemples regroupés en mini-lots peuvent être présentés. Un gradient moyen des erreurs est par la suite calculé pour indiquer la direction à prendre pour faire décroître le risque associé. Dans le cas des algorithmes de représentations où de nombreuses couches cachées sont présentes, la mise à jour des paramètres se fait en "rétro-propageant" le gradient de la couche de sortie vers la couche d'entrée par la règle de dérivées en chaîne.

$$\theta = \theta - \epsilon \nabla f(\theta) \tag{2.17}$$

où $f(\theta)$ correspond à la fonction apprise pour minimiser risque empirique et $\epsilon > 0$ au taux d'apprentissage. Une itération correspond à un passage à travers tout le jeu de données d'entrainement $\mathcal{D}_{train}$. En d'autres mots, une itération comprend de nombreuses mises à jour des paramètres.

Initialisation des paramètres

La descente de gradient stochastique rend les réseaux de neurones sensibles aux valeurs initiales des paramètres. Si ces valeurs s'éloignent d'une solution convenable, alors entrainer un ANN demeure très difficile (Hinton and Salakhutdinov, 2006). Une saturation trop importante des fonctions d'activation engendre une mauvaise propagation des gradients (Glorot and Bengio, 2010). De plus, les paramètres appris peuvent se révéler être une fonction identité ce qui rend l'apprentissage inutile (Glorot and Bengio, 2010).

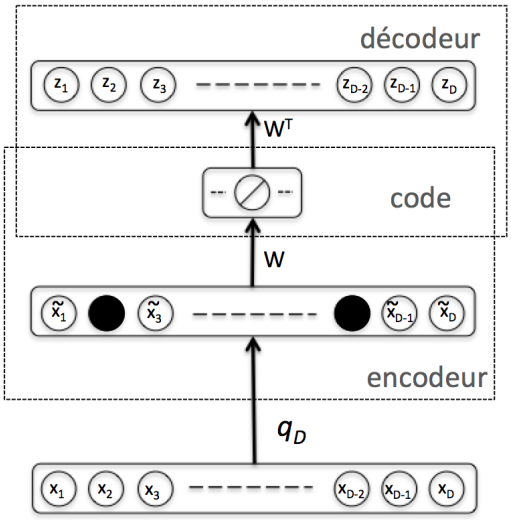


Figure 2.13 Auto-encodeur débruitant. Cette figure ressemble à la Figure 2.10 mis à par le fait qu'un auto-encodeur débruitant comporte un processus de corruption des entrées.

L'approche qui a permis l'essor récent de l'apprentissage de représentations profondes se base sur un modèle qui sert de pré-entrainement. Ce modèle est conçu pour reconstruire les données en entrée à partir d'une version corrompue. Des machines de Boltzmann restrictives (Hinton and Salakhutdinov, 2006) ou des auto-encodeurs débruitants (Vincent et al., 2008) peuvent être utilisés à cette fin. Nous nous attèlerons dans ce chapitre essentiellement à décrire cette dernière méthode, illustrée à la Figure 2.13, qui servira à initialiser les auto-encodeurs empilés du chapitre 5. Les auto-encodeurs débruitants peuvent être considérés comme une version stochastique des auto-encodeurs traditionnels (Vincent et al., 2008). La différence provient du processus de corruption stochastique, ou bruitage, qui impose aléatoirement certaines dimensions des exemples $x^{(i)}$ d'être égales à zéro. Cette version corrompue $\tilde{x}^{(i)}$ des entrées $x^{(i)}$ est obtenue par une transformation stochastique $\tilde{x} \sim q_D(\tilde{x}|x)$ avec une proportion de corruption υ. L'auto-encodeur débruitant est par la suite entrainé à reconstruire la version non-corrompue des entrées à partir d'une version corrompue. En d'autres mots, ce processus d'apprentissage tente d'annuler le processus de corruption stochastique en capturant les dépendances statistiques entre les entrées du jeu de données $\mathcal{D}$ (Bengio et al., 2013). Une fois le pré-entrainement de chacune des couches cachées effectuées, le réseaux de neurones peut procéder au raffinement des paramètres selon l'objectif d'apprentissage défini.

Bien que populaire lors de la (re-)naissance des algorithmes d'apprentissage de représentations, ces méthodes de pré-entrainement non-supervisé tendent de plus en plus à être occultés dans les articles de recherche les plus récents. La raison principale étant que cet apprentissage non-supervisé peut en réalité apprendre de mauvais paramètres car il est difficile d'en

contrôler la capacité (Bengio et al., 2015). La tendance va donc vers des méthodes d'apprentissage supervisé, où il est plus aisé de contrôler ce que l'algorithme apprend. Cela provient notamment des unités rectificatrices linéaires (section 2.6.2) et de la régularisation par dropout (section 2.6.3) en conjonction avec la grande disponibilité de données étiquetées qui permettent aujourd'hui d'entrainer des architectures à plusieurs couches cachées sans la nécessité d'algorithmes de pré-entrainement.

D'autres heuristiques plus simples qui ne nécessitent pas de pré-entrainement ont par conséquent pu être adoptées avec succès comme une initialisation par une distribution gaussienne (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Toutefois, l'entrainement d'architectures très profondes peut toujours se révéler difficile lorsque le nombre de couches cachées dépasse la dizaine (Simonyan and Zisserman, 2014).

Une distribution uniforme avec une moyenne de zéro et une variance unitaire peut également être utilisée (LeCun et al., 2012) :

$$W_{i,j} \sim U\Big[-\frac{1}{\sqrt{n_j}}, \frac{1}{\sqrt{n_j}}\Big] \tag{2.18}$$

où $U[-a, a]$ est une distribution uniforme portant sur l'intervalle $[-a, a]$, et n_j correspond au nombre d'unités de la couche précédente. Glorot and Bengio (2010) ont néanmoins montré que l'initialisation des poids de chacune des couches du réseau par cette même heuristique peut conduire à un gradient dont les valeurs s'évanouissent ou explosent pour un régime linéaire. Pour contrecarrer cet effet à travers les couches du réseau, l'idée est alors de normaliser l'initialisation des poids selon la formule suivante :

$$W_{i,j} \sim U\Big[-\frac{\sqrt{6}}{\sqrt{n_i + n_j}}, \frac{\sqrt{6}}{\sqrt{n_i + n_j}}\Big] \tag{2.19}$$

où n_i correspond au nombre d'unités de la couche courante et n_j à ceux de la couche précédente. Cette initialisation a pour effet de rendre la moyenne des valeurs singulières de la matrice Jacobienne des matrices de poids de chacune des couches proches de 1. Cette propriété démontre empiriquement une meilleure propagation des gradients (Glorot and Bengio, 2010).

Momentum

Le momentum est une technique pour accélérer la descente de gradient qui consiste à calculer une moyenne mobile des gradients à travers le temps. Concrètement dans le cas des ANNs,

cela consiste à accumuler un vecteur de vitesse v dans les directions qui permettent de réduire l'erreur estimée par la fonction objectif (Sutskever et al., 2013). La règle de mise à jour des paramètres de l'Éq. 2.17 se transforme alors en :

$$v_{t+1} = \mu v_t - \epsilon \nabla f(\theta_t) \tag{2.20}$$

$$\theta_{t+1} = \theta_t + v_{t+1} \tag{2.21}$$

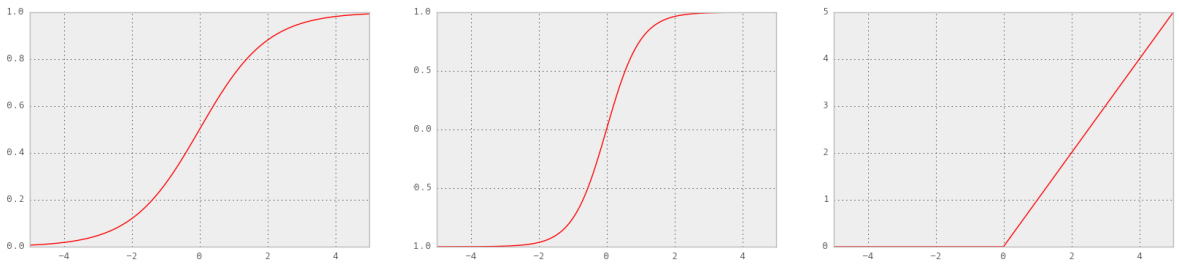
où $\mu \in [0, 1]$ représente le coefficient de momentum et ϵ le taux d'apprentissage. Le momentum de Nesterov est une variante du momentum classique qui consiste à mettre à jour les paramètres de la façon suivante :

$$v_{t+1} = \mu v_t - \epsilon \nabla f(\theta_t + \mu v_t) \tag{2.22}$$

$$\theta_{t+1} = \theta_t + v_{t+1} \tag{2.23}$$

La différence entre le momentum classique (Éq. 2.21) et le momentum de Nesterov (Éq. 2.23) réside dans le calcul du gradient. Le momentum de Nesterov effectue une première mise à jour avant le calcul du gradient, ce qui a pour effet de rendre le vecteur de vitesse v plus stable évitant les larges oscillations souvent observées durant l'apprentissage avec le momentum classique (Sutskever et al., 2013).

2.6.2 Fonctions d'activation



(a) Fonction logistique (b) Tangente hyperbolique (c) ReLU

Figure 2.14 Fonctions d'activation non-linéaires couramment utilisées.

La fonction d'activation la plus simple est l'activation *linéaire*, $s(x) = x$, qui retourne tout simplement les valeurs en sortie de la couche sans aucune modification. Cette activation s'utilise le plus souvent dans le cadre des tâches de régression où la couche de sortie doit produire une valeur réelle, mais aussi dans la couche du milieu d'un auto-encodeur empilé.

Dans la littérature d'autres fonctions d'activation existent également pour que les ANNs apprennent des fonctions non-linéaires. Les fonctions *sigmoïdes* – telles que la fonction *logistique*, $s(x) = 1/(1 + \exp(-x))$, illustrée à la Figure 2.14a, ou la *tangente hyperbolique*, $s(x) = \tanh(x)$, illustrée à la Figure 2.14b – sont parmi les plus communes. La *tangente hyperbolique* tend à être privilégiée par rapport à la fonction *logistique* car elle est symétrique par rapport à l'origine. Cette propriété est désirable car la *tangente hyperbolique* aura tendance à produire des valeurs d'activation (c'est-à-dire les valeurs à la sortie de la couche) dont la moyenne est de zéro. Ce comportement évite à l'entrainement de ralentir à cause de mises à jour des paramètres qui iraient dans une mauvaise direction (LeCun et al., 2012).

Aujourd'hui, la fonction *rectificatrice linéaire* ou *unité rectificatrice linéaire* (ReLU), est devenue la norme pour l'entrainement des ANNs. La forme particulière de la ReLU, illustrée à Figure 2.14c, incite l'ANN à apprendre des représentations parcimonieuses qui offrent de nombreux avantages (Glorot et al., 2011). La parcimonie pousse les ANN à apprendre à démêler les principaux facteurs de variation contrairement aux représentations denses, à modéliser la dimension effective des données, à mieux séparer les données entre elles, ou encore à distribuer l'information à travers les neurones artificiels. Empiriquement, l'utilisation des ReLUs a montré des résultats supérieurs aux méthodes utilisant un pré-entrainement (voir section 2.6.1), notamment parce que le réseau n'est plus sensible aux gradients qui s'évanouissent, effet présent avec les fonctions sigmoïdes (Glorot et al., 2011).

2.6.3 Régularisation

Afin de prévenir le sur-apprentissage des ANNs, plusieurs méthodes de régularisation des paramètres du modèle peuvent être appliquées lors de l'entrainement pour réduire l'erreur de généralisation sans affecter l'erreur d'entrainement (Bengio et al., 2015).

Norme des paramètres

La capacité des ANNs peut être restreinte en ajoutant la norme des paramètres à la fonction objectif. Cela se traduit soit par une régularisation de la norme L1 ($\lambda_1\|\theta\|_1 = \lambda_1 \sum_{d=1}^{D} |\theta_d|$) ou la norme L2 ($\lambda_2\|\theta\|_2 = \lambda_2 \sum_{d=1}^{D} \theta_d^2$). λ_1 et λ_2 sont des coefficients positifs qui modulent l'importance de chacune des régularisations. Si un des coefficients est égal à zéro, alors la régularisation correspondante n'est pas appliquée. La régularisation L1 encourage les paramètres non-discriminants à être égaux à zéro, ce qui crée de la parcimonie au sein du réseau (Bengio, 2012) et peut alors servir de sélection de traits caractéristiques (Ng, 2004). La régularisation L2 pénalise les paramètres dont la valeur est trop importante (Bengio, 2012), ce qui évite notamment des erreurs numériques causées par un dépassement en virgule flottante.

Arrêt prématuré

Le nombre de mises à jour des paramètres joue un grand rôle sur la capacité du modèle. Si le modèle n'est pas assez entrainé, cela conduit à un sous-apprentissage. Au contraire, si le modèle est entrainé sur une durée trop longue, cela conduit à un sur-apprentissage. Un compromis doit être trouvé pour optimiser la performance de généralisation. Une heuristique d'arrêt prématuré consiste à fixer les autres hyper-paramètres et à mesurer l'erreur du modèle sur le jeu de validation $\mathcal{D}_{valid}$ tout au long de l'entrainement et à l'arrêter lorsque cette erreur se dégrade. Cette heuristique peut être étendue pour éviter de s'arrêter trop tôt (Bengio, 2012). L'idée est de spécifier un nombre minimum, appelé *patience*, de mises à jour des paramètres à observer lorsqu'un minimum est atteint avec $\mathcal{D}_{valid}$. Si un nouveau minimum est découvert durant la *patience* établie, alors sa longueur est étendue. Si aucun nouveau minimum n'est découvert, alors l'entrainement s'arrête et les paramètres reviennent à l'état du dernier minimum atteint. Il est à noter que l'erreur sur $\mathcal{D}_{valid}$ ne doit pas être estimée après chacune mise à jour des paramètres, mais à chaque itération pour éviter de conduire à des temps de calcul trop important liés à cette heuristique d'arrêt prématuré (Bengio, 2012).

Ensembles

L'optimisation d'un ANN étant non-convexe, de nombreux minima locaux sont présents. Un minimum local se définit comme un point dans l'espace des paramètres qui conduit à une erreur plus faible que celle de ses points voisins (Bengio et al., 2015). En faisant des petits pas de gradients, il demeure donc difficile d'en sortir. Dans le cas de l'apprentissage de représentations, ces minima locaux permettent toutefois d'obtenir des performances remarquables. Le défi repose plutôt sur la compréhension des points de selle qui sont plus dommageables pour l'optimisation des ANNs (Bengio et al., 2015). La présence de nombreux minima locaux signifie que selon les valeurs initiales des paramètres, les résultats finaux du modèle appris diffèreront (Hansen and Salamon, 1990). En d'autres mots, ces modèles feront des erreurs lors de la prédiction, mais pas forcément les mêmes erreurs. L'idée des ensembles est alors de tirer partie de cette propriété pour créer une prédiction collective car un ensemble tend à produire une meilleure prédiction qu'un modèle pris individuellement (Hansen and Salamon, 1990). Pour créer des ensembles, plusieurs heuristiques existent. Chaque modèle peut être entrainé avec des initialisations différentes, ou encore des mini-lots présentés dans un ordre différent (Hansen and Salamon, 1990). La décision collective quant à elle peut consister en des votes majoritaires (la prédiction finale est celle ayant atteint au moins 50% des voix), ou bien des votes pluraux (la prédiction finale est celle ayant reçu le plus de voix). Dans la plupart des compétitions d'apprentissage automatique, telles que ImageNet (Russakovsky

et al., 2014), l'utilisation d'ensembles de modèles est devenue la norme pour atteindre les meilleures performance de généralisation.

Dropout

La régularisation par dropout est un procédé de corruption qui consiste appliquer une transformation stochastique pour imposer certaines dimensions de la sortie d'une couche à être égales à zéro (Srivastava et al., 2014). En ce sens, cette transformation stochastique ressemble au procédé de corruption d'un auto-encodeur débruitant. La différence réside dans l'utilisation de la corruption. Lors de l'entrainement, la régularisation par dropout ne retient qu'un certain nombre d'unités selon une proportion p, ce qui résulte en un ANN de plus petite taille. Seuls les paramètres des unités retenues sont mises à jour. En d'autres mots, pour chaque mise à jour des paramètres, un réseau différent est échantillonné. Lors de la phase de test, cet ensemble exponentiel d'ANNs différents de petite taille est approximé en multipliant les paramètres des unités cachées par leur probabilité de rétention p. La régularisation par dropout offre alors des avantages similaires aux ensembles (section 2.6.3) mais dans une moindre mesure.

Un des avantages majeurs de la régularisation par dropout est qu'elle empêche la co-adaptation entre les unités cachées afin de favoriser l'apprentissage de traits caractéristiques discriminants (Srivastava et al., 2014). En effet, il se peut qu'une unité corrige les erreurs effectuées par son unité voisine, ce qui engendre des effets de co-adaptation complexes entre les unités cachées. Ces effets sont néfastes pour l'entrainement d'un ANN car ils conduisent à un sur-apprentissage. De plus, la régularisation par dropout encourage l'apprentissage de représentations parcimonieuses (Srivastava et al., 2014) qui présentent de nombreux avantages pour faciliter l'entrainement d'un ANN (voir section 2.6.2 sur les ReLUs).

Cependant, la régularisation par dropout requiert un temps d'entrainement 2 à 3 fois plus important qu'un entrainement classique sans dropout (Srivastava et al., 2014). Ceci provient du fait que chaque mise à jour des paramètres est calculée à partir d'une nouvelle architecture échantillonnée aléatoirement. Il y a donc un compromis à adopter entre le temps alloué à l'entrainement et le sur-apprentissage du modèle.

Augmentation artificielle de données

Une simple heuristique pour prévenir le sur-apprentissage consiste à augmenter artificiellement les données. Concrètement, cela signifie que lors de l'entrainement, des transformations aléatoires vont être appliquées à chacun des exemples présentés au ANN. Étant donné la na-

ture aléatoire, il se peut qu'aucune transformation ne soit appliquée, que l'exemple ne soit que très peu modifié, ou qu'il ait été fortement modifié. Parmi les transformations couramment utilisées, on retrouve les transformations affines qui regroupent notamment les translations, les mises à l'échelle, les réflexions, les rotations ou encore les transvections (similaires à une contrainte de cisaillement). D'autres transformations plus avancées peuvent par exemple altérer les intensités des canaux couleurs de l'image en ajoutant du bruit aux composantes principales de l'image (Krizhevsky et al., 2012).

2.6.4 Choix des hyper-paramètres

La recherche d'hyper-paramètres dans le domaine de l'apprentissage automatique, et plus spécialement l'apprentissage de représentations, est parfois considérée comme un art. Le nombre d'hyper-paramètres présents est en effet très conséquent et choisir les valeurs engendrant des performances élevées n'est pas trivial. Comme mentionné en introduction, cette recherche se fait sur le jeu de validation $\mathcal{D}_{valid}$. Certaines valeurs, ou plutôt certaines gammes de valeurs, d'hyper-paramètres existent de par des résultats empiriques passés. Ces valeurs ne sont toutefois pas définies dans le marbre puisque pour chaque problème, et pour chaque jeu de données, correspond un ensemble d'hyper-paramètres différents (Bengio, 2012).

L'heuristique la plus simple consiste à faire une recherche manuelle. Les algorithmes d'apprentissage automatique sont alors très dépendants de l'interaction avec l'utilisateur qui doit spécifier la valeur de chacun des hyper-paramètres. Une autre heuristique consiste à spécifier une grille de la dimension du nombre d'hyper-paramètres que l'on souhaite déterminer. L'idée est d'entrainer un modèle pour toutes les combinaisons possibles de la grille et de choisir l'ensemble d'hyper-paramètres produisant l'erreur la plus faible sur le jeu de validation.

La recherche d'hyper-paramètres sur une grille est néanmoins très exhaustive, si l'on sait qu'une zone de la grille n'engendre pas de bons résultats, pourquoi continuer à essayer les hyper-paramètres de cette zone ? Une heuristique différente consiste à échantillonner les valeurs des hyper-paramètres selon une loi uniforme (Bergstra and Bengio, 2012). La recherche des hyper-paramètres se fait alors de manière aléatoire et couvre alors des zones non exploitées par une recherche sur une grille. Cette heuristique peut être étendue pour tirer avantage de toutes les décisions passées. L'idée est d'utiliser l'optimisation bayésienne (Snoek et al., 2012) qui va décider de l'espace à explorer pour l'itération suivante selon les évaluations passées qui servent d'a priori.

CHAPITRE 3 MÉTHODOLOGIE

3.1 Classification de la scoliose idiopathique de l'adolescent

Les chapitres 4 et 5 présentent l'utilisation d'auto-encodeurs empilés (section 2.5.3) pour l'apprentissage de représentations de manière non-supervisée de modèles géométriques correspondant à la reconstruction de la colonne vertébrale en trois dimensions de patients atteints de scoliose idiopathique adolescente.

Les colonnes vertébrales des patients sont représentées par des modèles géométriques qui consistent en des coordonnées dans un espace normalisé de marqueurs placés à des positions spécifiques sur les vertèbres de la colonne vertébrale. L'ensemble des marqueurs constitue une reconstruction en trois dimensions de la colonne vertébrale. Les marqueurs ont été générés de manière semi-automatique à partir de radiographies biplanaires issues de systèmes classiques (section 2.2.1) ou de systèmes EOS (section 2.2.2).

L'apprentissage de représentations de manière non-supervisée sert alors à apprendre une représentation plus compacte des reconstructions du rachis pour faciliter le démêlement des principaux facteurs de variations. Un autre algorithme d'apprentissage automatique non-supervisé, les k-moyennes++, partitionne par la suite les représentations de faibles dimensions en sous-groupes cliniquement significatifs qui serviront à établir une classification des patients.

Dans un premier temps, la méthodologie proposée a été appliquée à une base de donnée restreinte, composée de 277 reconstructions issues de 155 patients de type Lenke I. En a découlé une première présentation orale à l'atelier Computational Methods and Clinical Applications for Spine Imaging (CSI) de la conférence Medical Image Computing and Computer Assisted Intervention (MICCAI) 2014 qui constitue le chapitre 4 :

— **W. Thong**, H. Labelle, J. Shen, S. Parent, et S. Kadoury, "Stacked Auto-Encoders for Classification of 3D Spine Models in Adolescent Idiopathic Scoliosis", *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pp. 13-25, 2014.

Les travaux de cet article ont par la suite été étendus à une base de données multicentriques, de neuf centres de recherche sur la scoliose à travers l'Amérique du Nord, et composée de 915 reconstructions issues de 663 patients représentants tous les types de Lenke. En a découlé une soumission à *European Spine Journal* qui constitue le chapitre 5 :

— **W. Thong**, S. Parent, J. Wu, C.-E. Aubin, H. Labelle, et S. Kadoury, "Three-

Dimensional Classification of Adolescent Idiopathic Scoliosis from Encoded Geometric Models", *European Spine Journal*, soumis, 2015.

3.2 Classification de voxels pour la segmentation

Le chapitre 6 présente l'utilisation de réseaux à convolution (section 2.5.4) pour l'apprentissage de représentations de manière supervisée d'images médicales. L'objectif d'apprentissage consiste à segmenter les reins dans les images tomodensitométriques abdominales.

Le problème de segmentation a été reformulé en un problème de classification. Le réseau à convolution apprend à classifier chaque voxel de l'image selon son appartenance à l'avant-plan (c'est-à-dire les reins) ou à l'arrière-plan (c'est-à-dire le reste du corps humain). La segmentation des reins dans un volume d'images revient alors à classifier tous les voxels présents dans l'image.

L'entrainement se fait avec des petits blocs de l'image, appelés *patches*, échantillonnés aléatoirement au sein du jeu de données. Les deux classes comprennent le même nombre de *patches* pour éviter que le ConvNet n'apprenne une fonction biaisée envers une classe. Pour obtenir la segmentation des reins de simples modifications de l'architecture du réseau à convolution sont opérées pour que le réseau puisse prendre en compte des images médicales entièrement au lieu des patches de plus petite taille.

En a découlé une présentation orale à l'atelier Deep Learning in Medical Image Analysis (DLMIA) de la conférence MICCAI 2015 qui constitue le chapitre 6 :

— **W. Thong**, S. Kadoury, N. Piché, C.J. Pal, "Convolutional Networks for Kidney Segmentation in Contrast-Enhanced CT Scans", *Lecture Notes in Computational Vision and Biomechanics*, Springer International Publishing, (sous presse), 2015

CHAPITRE 4 ARTICLE #1 : STACKED AUTO-ENCODERS FOR CLASSIFICATION OF 3D SPINE MODELS IN ADOLESCENT IDIOPATHIC SCOLIOSIS

Cet article présente l'utilisation de réseaux de neurones pour l'apprentissage de représentations de modèles géométriques en trois dimensions correspondant à la colonne vertébrale de patients atteints de scoliose idiopathique adolescente. Cet article pose les bases méthodologiques pour l'article présenté au Chapitre 5. L'article au sein de ce chapitre ne porte toutefois que sur des déformations thoraciques classées Lenke Type-1 (155 patients pour 277 reconstructions).

Auteurs

William Thong[a,b], Hubert Labelle[b], Jesse Shen[b], Stefan Parent[b], Samuel Kadoury[a,b]

Affiliations

[a] Polytechnique Montréal, Montréal, Québec, Canada

[b] CHU Sainte-Justine, Montréal, Québec, Canada.

4.1 Abstract

Current classification systems for adolescent idiopathic scoliosis lack information on how the spine is deformed in three dimensions (3D), which can mislead further treatment recommendations. We propose an approach to address this issue by a deep learning method for the classification of 3D spine reconstructions of patients. A low-dimensional manifold representation of the spine models was learnt by stacked auto-encoders. A K-Means++ algorithm using a probabilistic seeding method clustered the low-dimensional codes to discover sub-groups in the studied population. We evaluated the method with a case series analysis of 155 patients with Lenke Type-1 thoracic spinal deformations recruited at our institution. The clustering algorithm proposed 5 sub-groups from the thoracic population, yielding statistically significant differences in clinical geometric indices between all clusters. These results demonstrate the presence of 3D variability within a pre-defined 2D group of spinal deformities.

4.2 Introduction

Adolescent idiopathic scoliosis (AIS) refers to a complex deformation in three dimensions (3D) of the spine. Classification of the rich and complex variability of spinal deformities is critical for comparisons between treatments and for long term patient follow-ups. AIS characterization and treatment recommendations currently rely on the Lenke classification system (Lenke et al., 2001) because of its simplicity and its high inter- and intra-observer reliability compared with previous classification systems (King et al., 1983). However, these schemes are restricted to a two-dimensional (2D) assessment of scoliosis from radiographs of the spine in the coronal and sagittal plane. Misinterpretations could arise because two different scoliosis deformities may have similar 2D parameters. Therefore, improvements in the scoliosis classification system are necessary to ensure a better understanding and description of the curve morphology.

Computational methods open up new paths to go beyond the Lenke classification. Recent studies seek new groups in the population of AIS using cluster analysis (Duong et al., 2006, 2009; Sangole et al., 2009; Stokes et al., 2009) with ISOData, K-Means or fuzzy C-Means algorithms. Their common aspect is founded upon the clustering of expert-based features, which are extracted from 3D spine reconstructions (Cobb angles, kyphosis and planes of maximal deformity). This methodology stems from the fact that clustering algorithms are very sensitive to the curse of dimensionality. Still, these parameters might not be enough to tap all the rich and complex variability in the data. Computational methods should be able to capture the intrinsic dimensionality that explain as much as possible the highly dimensional

data into a manifold of much lower dimensionality (Bengio et al., 2013; van der Maaten et al., 2009). Hence, another paradigm for spine classification is to let the algorithm learn its own features to discriminate between different pathological groups. This implies directly analyzing the 3D spine models instead of expert-based features as it has been experimented previously. To our knowledge, only one study tried to learn a manifold from the 3D spine model (Kadoury and Labelle, 2012) using Local Linear Embeddings (LLE). In this study, we propose to use stacked auto-encoders –a deep learning algorithm– to reduce the high-dimensionality of 3D spine models in order to identify particular classes within Lenke Type-1 curves. This algorithm was able to outperform principal component analysis (PCA) and LLE (Hinton and Salakhutdinov, 2006; van der Maaten et al., 2009).

Recent breakthroughs in computer vision and speech processing using deep learning algorithms suggest that artificial neural networks might be better suited to learn representations of highly non-linear data (Bengio et al., 2013). Training a deep neural network has been a challenging task in many applications. Nowadays, this issue is tackled by leveraging more computation power (i.e. parallelizing tasks), more data and by a better initialization of the multilayer neural network (Hinton and Salakhutdinov, 2006). Deep neural networks promote the learning of more abstract representations that result in improved generalization. Network depth actually helps to become invariant to most local changes in the data and to disentangle the main factor of variations in the data (Bengio et al., 2013).

We propose a computational method for the classification of highly dimensional 3D spine models obtained from a group of patients with AIS. The core of the methodology, detailed in Section 4.3, builds a low-dimensional representation of the 3D spine model based on stacked auto-encoders that capture the main variabilities in the shape of the spine. The low-dimensional codes learnt by the stacked auto-encoders are then clustered using the K-Means++ algorithm. Finally, a statistical analysis assesses the relevance of the clusters identified by the framework based on clinical geometrical indices of the spine. Experiments conducted with this methodology are shown and discussed in Section 4.4, while Section 4.5 concludes this paper.

4.3 Methods

The proposed method consists of four main steps: (1) reconstruction of a 3D spine model from biplanar X-rays for each patient; (2) dimensionality reduction of each high-dimensional model to a low-dimensional space; (3) clustering of the low-dimensional space; (4) analysis of the clusters obtained with the clinical data.

4.3.1 3D spine reconstruction

A 3D model for each patient's spine was generated from anatomical landmarks with a semi-supervised statistical image-based technique built in a custom software in C++ (Pomero et al., 2004). Seventeen 3D anatomical landmarks were extracted per vertebra (12 thoracic, 5 lumbar): center, left, right, anterior and posterior of superior and inferior vertebral endplates (10 landmarks); left and right transverse process (2 landmarks); spinous process (1 landmark); and tips of both pedicles (4 landmarks). All 3D spine models were normalized with regards to their height and rigidly translated to a common referential at the L5 vertebra. Hence, each observation contains 867 features, which corresponds to the concatenation of the 3D coordinates of all the landmarks into an observation vector.

An auto-encoder is a neural network that learns a hidden representation to reconstruct its input. Consider a one hidden layer auto-encoder network. First, the input vector $\mathbf{x}$ of dimension d, representing the 3D coordinates of a spine model, is mapped by an encoder function f into the hidden layer $\mathbf{h}$, often called a code layer in the case of auto-encoders:

$$\mathbf{h} = f(\mathbf{x}) = s(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \tag{4.1}$$

where $\mathbf{W}^{(1)}$ is the encoding weight matrix, $\mathbf{b}^{(1)}$ the bias vector and $s(\cdot)$ the activation function. Note that this one hidden layer auto-encoder network corresponds to a principal component analysis if the activation function is linear. Then, the code representation is mapped back by a decoder function g into a reconstruction $\mathbf{z}$:

$$\mathbf{z} = g(f(\mathbf{x})) = s(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}) \tag{4.2}$$

where $\mathbf{W}^{(2)}$ is the decoding weight matrix. Tying the weights ($\mathbf{W}^{(2)} = \mathbf{W}^{(1)^T}$) has several advantages. It acts as a regularizer by preventing tiny gradients and it reduces the number of parameters to optimize (Bengio et al., 2013). Finally, the parameters $\theta = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}\}$ are optimized in order to minimize the squared reconstruction error:

$$L_2(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|^2 \ . \tag{4.3}$$

In the case of dimensionality reduction, the code layer $\mathbf{h}$ has a smaller dimension than the input $\mathbf{x}$. One major drawback comes from the gradient descent algorithm for the training procedure that is very sensitive to the initial weights. If they are far from a good solution, training a deep non-linear auto-encoder network is very hard (Hinton and Salakhutdinov, 2006). A pre-training algorithm is thus required to learn more robust features before fine-

tuning the whole model.

The idea for initialization is to build a model that reconstructs the input based on a corrupted version of itself. This can either be done by Restricted Boltzmann Machines (RBMs) (Hinton and Salakhutdinov, 2006) or denoising auto-encoders (Vincent et al., 2008) (used in this study). The denoising auto-encoder is considered as a stochastic version of the auto-encoder (Vincent et al., 2008). The difference lies in the stochastic corruption process that sets randomly some of the inputs to zero. This corrupted version $\tilde{\mathbf{x}}$ of the input $\mathbf{x}$ is obtained by a stochastic mapping $\tilde{\mathbf{x}} \sim q_D(\tilde{\mathbf{x}}|\mathbf{x})$ with a proportion of corruption υ. The denoising auto-encoder is then trained to reconstruct the uncorrupted version of the input from the corrupted version, which means that the loss function in equation 4.3 remains the same. Therefore, the learning process tries to cancel the stochastic corruption process by capturing the statistical dependencies between the inputs (Bengio et al., 2013). Once all the layers are pre-trained, the auto-encoder proceeds to a fine-tuning of the parameters θ (i.e. without the corruption process).

4.3.2 K-Means++ Clustering algorithm

Once the fine-tuning of the stacked auto-encoder has learnt the parameters θ, low-dimensional codes from the code layer can be extracted for each patient's spine. Clusters in the codes were obtained using the K-Means++ algorithm (Arthur and Vassilvitskii, 2007), which is a variant of the traditional K-Means clustering algorithm but with a selective seeding process. First, a cluster centroid is initialized among a random code layer $\mathbf{h}$ of the dataset χ following a uniform distribution. Afterwards, a probability is assigned to the rest of the observations for choosing the next centroid:

$$p(\mathbf{h}) = \frac{D(\mathbf{h})^2}{\sum_{h \in \chi} D(\mathbf{h})^2} \tag{4.4}$$

where $D(\mathbf{h})^2$ corresponds to the shortest distance from a point $\mathbf{h}$ to its closest cluster centroid. After the initialization of the cluster centroids, the K-Means++ algorithm proceeds to the regular Lloyd's optimization method.

The selection of the right number of clusters k is based on the validity ratio (Ray and Turi, 1999), which minimizes the intra-cluster distance and maximizes the inter-cluster distance. The ratio is defined as $validity = intra/inter$. The intra-cluster distance is the average of all the distances between a point and its cluster centroid:

$$\text{intra} = \frac{1}{N}\sum_{i=1}^{k}\sum_{\mathbf{x}\in C_i}\|\mathbf{h}-\mathbf{c}_i\|^2 \ . \tag{4.5}$$

where N is the number of observations in χ and $\mathbf{c}_i$ the centroid of cluster i. The inter-cluster distance is the minimum distance between cluster centroids.

$$\text{inter} = \min(\|\mathbf{c}_i - \mathbf{c}_j\|^2) \tag{4.6}$$

where $i = 1, 2, \cdots, k-1$ and $j = i+1, \cdots, k$.

Clinical data analysis

Once clusters were created from the low-dimensional representation of the dataset, we analyzed the clustered data points with 3D geometric indices in the main thoracic (MT) and thoracolumbar/lumbar (TLL) regions for each patient's spine. One-way ANOVA tested differences between the cluster groups with a significance level $\alpha = 0.05$. The p-values were adjusted with the Bonferroni correction. For all cases, the following 3D spinal indices were computed: the orientation of the plane of maximum curvature (PMC) in each regional curve, which corresponds to the plane orientation where the projected Cobb angle is maximal; the kyphotic angle, measured between T2 and T12 on the sagittal plane; the lumbar lordosis angle, defined between L1 and S1 on the sagittal plane; the Cobb angles in the MT and TLL segments; and the axial orientation of the apical vertebra in the MT region, measured by the Stokes method (Stokes et al., 1986).

4.4 Clinical experiments

Clinical data

A cohort of 155 AIS patients was recruited for this preliminary study at our institution. A total of 277 reconstructions of the spine was obtained in 3D. From this group, 60 patients had repeat measurements from multiple clinic visits (mean = 3 visits). The mean thoracic Cobb angle was $53.2 \pm 18.3^\circ$ (range = $11.2 - 100.2^\circ$). All patients were diagnosed with a right thoracic deformation and classified as Lenke Type-1 deformity. A lumbar spine modifier (A, B, C) was also assigned to each observation, using the biplanar X-Ray scans available for each patient. The dataset included 277 observations divided in 204 Lenke Type-1A, 43 Lenke Type-1B and 30 Lenke Type-1C deformities. The training set included 235 observations, while the validation set included 42 observations (15% of the whole dataset). Observations

were randomly assigned in each set.

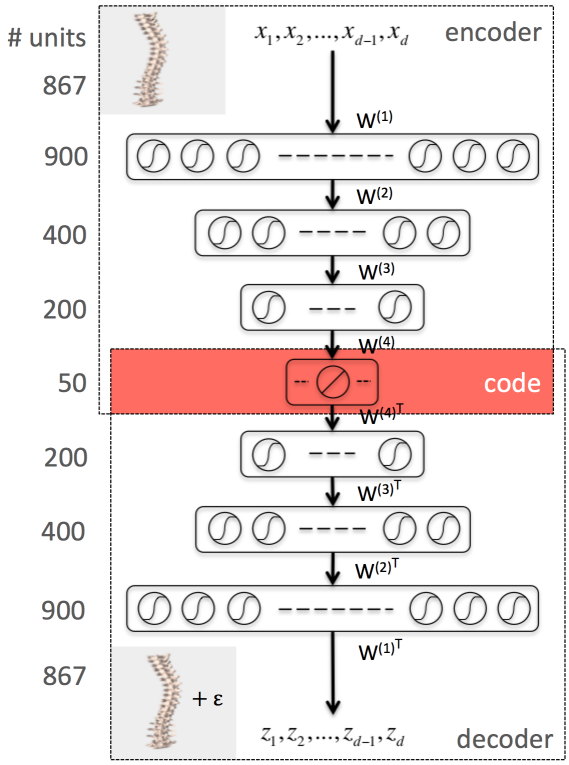


Figure 4.1 Illustration of the stacked auto-encoders architecture to learn the 3D spine model by minimizing the loss function. The middle layer represents a low-dimensional representation of the data, which is named the code layer. An optimal layer architecture of 867-900-400-200-50 was found after a coarse grid search of the hyper-parameters.

4.4.1 Hyper-parameters of the stacked auto-encoders

The neural network hyper-parameters were chosen by an exhaustive grid search. The architecture yielding to the lowest validation mean squared error (MSE) is described in Figure 4.1. We used an encoder with layers of size 867-900-400-200-50 and a decoder with tied weights to map the high-dimensional patient's spine models into low-dimensional codes. All units in the network were activated by a sigmoidal function $s(a) = 1/(1 + e^{-a})$, except for the 50 units in the code layer that remain linear $s(a) = a$.

4.4.2 Training and testing the stacked auto-encoders

Auto-encoder layers were pre-trained and fine-tuned with the stochastic gradient descent method using a GPU implementation based on the Theano library (Bergstra et al., 2010). Pre-training had a proportion of corruption for the stochastic mapping in each hidden layer of

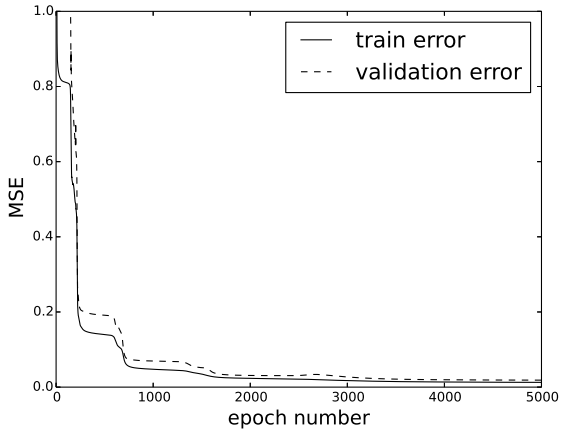


Figure 4.2 Evolution of the mean squared error (MSE) with respect to the number of epochs to determine the optimal model described in Fig. 4.1.

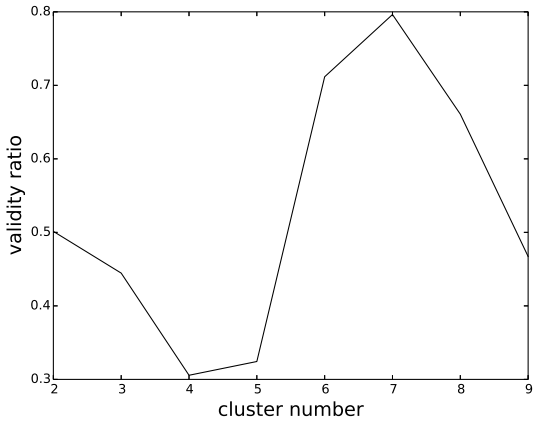
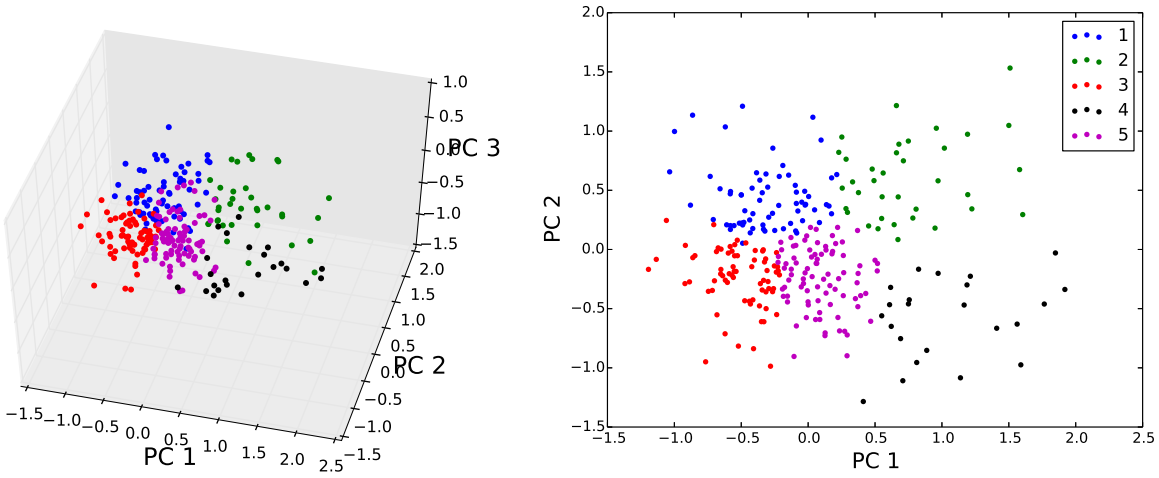


Figure 4.3 Validity ratio with respect to the number of clusters, to determine the optimal number of clusters.

$\upsilon = \{0.15, 0.20, 0.25, 0.30\}$ and a learning rate of $\epsilon_0 = 0.01$. Fine-tuning ran for 5000 epochs and had a learning rate schedule with $\epsilon_0 = 0.05$ that annihilates linearly after 1000 epochs. Figure 6.2 shows the learning curve of the stacked auto-encoder. The optimal parameters θ for the model in Figure 4.1 were found at the end of training with a normalized training MSE of 0.0127, and a normalized validation MSE of 0.0186, which corresponds to 4.79 mm^2 and 6.46 mm^2 respectively on the original scale. The learning curve describes several flat regions before stabilizing after 3500 epochs.

4.4.3 Clustering the codes

The K-Means++ algorithm was done using the scikit-learn library (Pedregosa et al., 2011). For each number of clusters k (2 through 9), the algorithm ran for 100 times with different centroid seeds in order to keep the best clustering in terms of inertia. Figure 4.3 depicts the validity ratio against the number of clusters. The validity ratio suggests that the optimal number of clusters should be 4 or 5. However, subsequent analysis illustrated in Table 4.1 indicates that 5 clusters is the right number of clusters for this dataset because all the clinical indices are statistically significant ($\alpha = 0.05$) given that the other indices are in the model. Figure 4.4 presents the visualization of the five clusters using a PCA to project the codes in 3D and 2D views. However, it should be mentioned that the clustering was performed on the codes of dimension 50. Figure 4.5 shows the frontal, lateral and daVinci representations of the centroid of each cluster.



(a) Three-dimensional view

(b) Two-dimensional view

Figure 4.4 Visualization of the five clusters found by the K-Means++ algorithm on low-dimensional points, by projecting the 50-dimensional codes into 3 (a) and 2 (b) principal components (PC) using PCA.

4.4.4 Clinical significance

Based on the five identified clusters, cluster 1 is composed of 50 Lenke Type-1A, 12 Type-1B and 2 Type-1C, representing hyper-kyphotic, hyper-lordotic profiles, with high curvatures in the sagittal plane. No sagittal rotation was detected in cluster 1. Cluster 2 is composed of 29 Lenke Type-1A, 7 Type-1B and 0 Type-1C, representing a high axial rotation of the apical vertebra, with the strongest thoracic deformation of all clusters. Moreover, those two clusters have no lumbar derotation.

Clusters 3, 4 and 5 represent the clusters with higher lumbar deformities. Cluster 3 includes 34 Lenke Type-1A, 17 Type-1B and 20 Type-1C, with a minimal thoracic deformation and the highest angulation of TLL plane of all clusters. Cluster 4 includes 23 observations, with 21 Lenke Type-1A, 2 Type-1B and 0 Type-1C. Cluster 4 is characterized by a hypo-kyphotic profile (mean = 7°) and the highest angulation of the MT plane of all clusters. Finally, cluster 5 includes 70 Lenke Type-1A, 5 Type-1B and 8 Type-1C, with a low kyphosis, and medium range thoracic deformations. While this last cluster has a higher orientation of the thoracolumbar curve, its magnitude is not significant.

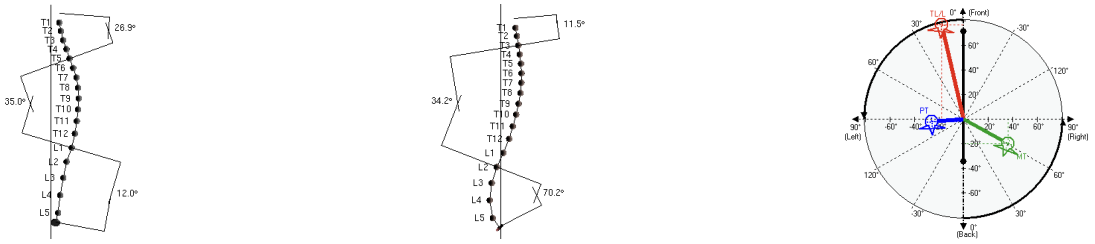
Surgical strategies based on current 2D classification systems are suboptimal since they do not capture the intrinsic 3D representation of the deformation. Lenke Type-1 classification currently leads to selective thoracic arthrodesis. This very restrictive surgery treatment

Table 4.1 Mean geometric clinical 3D parameters for the thoracic and lumbar regions, within all five clusters detected by the framework.

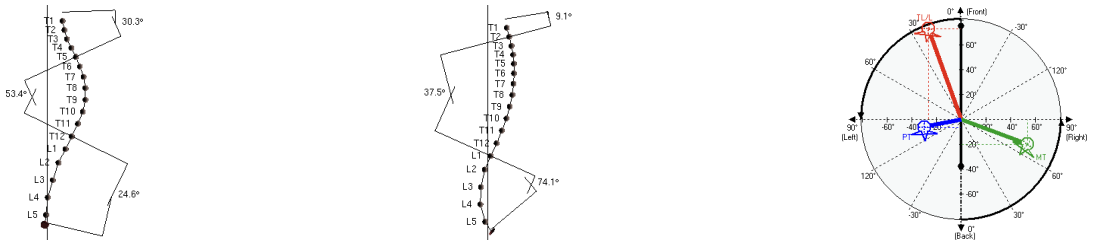
Parameter (all in degrees)	**Cluster 1** (n=64)	**Cluster 2** (n=36)	**Cluster 3** (n=71)	**Cluster 4** (n=23)	**Cluster 5** (n=83)	p-value$_{\text{ajd}}$
PMC MT Cobb	50.6±13.8	73.3±12.7	43.3±15.5	62.2±9.2	52.5±14.6	<0.001 *
PMC MT Orient	71.6±12.1	78.1±7.5	72.5±12.1	88.5±4.1	82.0±10.4	<0.001 *
PMC TLL Cobb	34.1±12.7	49.5±11.6	36.2±15.1	38.2±12.1	36.6±13.3	<0.001 *
PMC TLL Orient	45.4±18.9	51.3±15.0	65.0±24.0	52.8±12.2	48.4±17.3	<0.001 *
Kyphosis	36.6±10.6	32.4±11.2	27.4±9.8	7.3±11.6	21.9±11.7	<0.001 *
Lordosis	-66.6±8.9	-63.4±14.1	-61.6±12.3	-54.2±13.1	-60.9±9.7	0.002 *
MT Cobb	46.6±22.2	69.2±24.8	43.3±15.6	62.2±9.2	50.9±18.4	<0.001 *
TLL Cobb	-29.9±22.2	-45.5±25.5	-36.3±14.9	-38.2±12.1	-34.8±18.3	0.036 *
Ax. rot. apex	-19.7±10.3	-33.8±7.6	-13.7±8.7	-25.9±8.8	-22.1±12.0	<0.001 *

A significant star (*) indicates that the differences of geometric parameter between all groups are statistically significant at level $\alpha = 0.05$, given that the other geometric parameters are in the model.

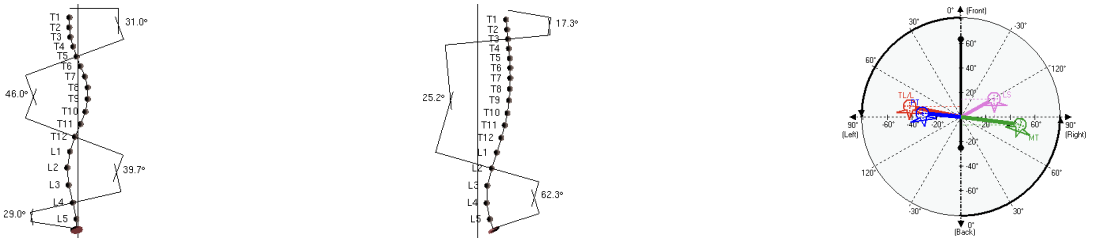
comes from the hard thresholds on the geometric parameters. A small change in Cobb angle could lead to two different classification and to two different fusion recommendations subsequently (Labelle et al., 2011). Therefore, identifying groups based on their true 3D representation will help to better adjust surgery choices such as levels of fusion, biomechanical forces to apply or surgical instrumentations. In this study, the learning framework provided an optimal number of 5 clusters based on the input population. It is not possible at this stage to infer that Lenke Type-1 should be divided in 5 groups. However, this study confirms that within a defined 2D class currently used for surgical planning, there exists a number of sub-groups with different 3D signatures that are statistically significant. Therefore, each subgroup would lead to different surgical strategies. Previous studies have indicated that within Lenke Type-1 (Duong et al., 2006; Sangole et al., 2009; Kadoury and Labelle, 2012), there indeed exists 3D variability in terms of geometric parameters that could be divided in 4 to 6 sub-groups.



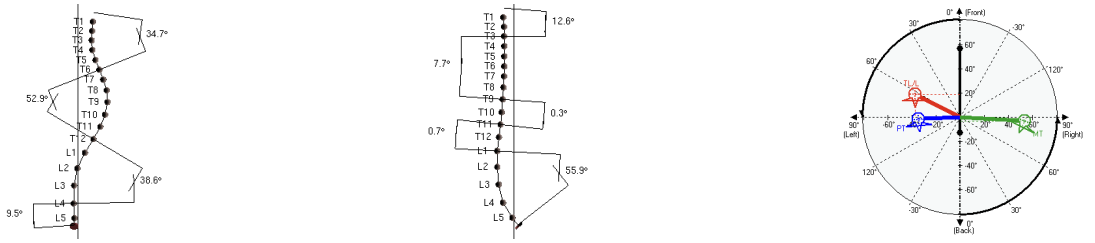
(a) Centroid of cluster 1



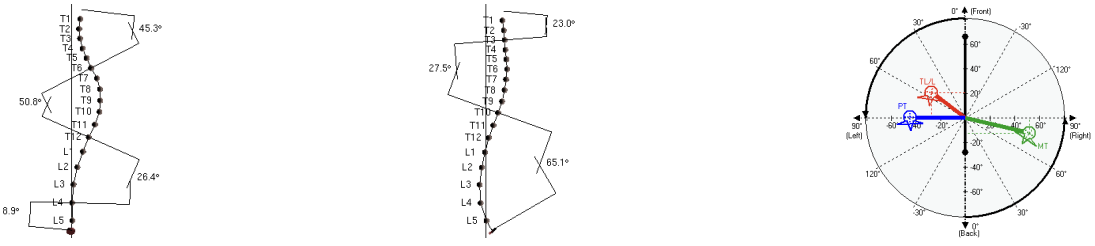
(b) Centroid of cluster 2



(c) Centroid of cluster 3



(d) Centroid of cluster 4



(e) Centroid of cluster 5

Figure 4.5 Frontal, lateral and top view profiles of cluster centers, with daVinci representations depicting planes of maximal deformities.

4.5 Conclusion

In this paper, we presented an automated classification method using a deep learning technique, namely with stacked auto-encoders, to discover sub-groups within a cohort of thoracic deformations. The code layer of the auto-encoder learns a distributed representation in low-dimension that aims to capture the main factors of variation in the training dataset. However, different examples from the distribution of the training dataset may potentially yield to high reconstruction errors (Bengio et al., 2013). Therefore, having a large and representative training dataset of AIS is critical. This will also prevent the model from overfitting.

The current study evaluated the 3D sub-classification of Lenke Type-1 scoliotic curves, suggesting that shape variability is present within an existing 2D group used in clinical practice. However, these types of approaches include complex synthetization tasks, which require sizeable datasets to improve the data representation within the code layer. Therefore, a multicentric dataset may help to significantly increase the number of cases from various sites and obtain a more reproducible model. Furthermore, the development of computational methods will ultimately lead to more reliable classification paradigms, helping to identify possible cases which might progress with time. Future work will include additional Lenke types, such as double major and lumbar deformations. Other works will propose to use longitudinal data for surgical treatment planning, whereas each observation is considered independently in the current framework. Finally, a reliability study will be undertaken to evaluate the relevance of classification systems.

4.6 Acknowledgements

This paper was supported in part by the CHU Sainte-Justine Academic Research Chair in Spinal Deformities, the Canada Research Chair in Medical Imaging and Assisted Interventions, the 3D committee of the Scoliosis Research Society, the Natural Sciences and Engineering Research Council of Canada and the MEDITIS training program.

CHAPITRE 5 ARTICLE #2 : THREE-DIMENSIONAL CLASSIFICATION OF ADOLESCENT IDIOPATHIC SCOLIOSIS FROM ENCODED GEOMETRIC MODELS

Cet article présente l'utilisation de réseaux de neurones pour l'apprentissage non-supervisé de représentations de modèles géométriques en trois dimensions correspondant à la colonne vertébrale de patients atteints de scoliose idiopathique adolescente. Il fait suite à l'article du précédant chapitre mais avec une large banque de patients (663 patients pour 915 reconstructions) pour couvrir au mieux les déformations de la colonne vertébrale observées par les chirurgiens.

Authors

William Thong[a,b], Stefan Parent[b], James Wu[b], Carl-Eric Aubin[a,b], Hubert Labelle[b], Samuel Kadoury[a,b]

Affiliations

[a] Polytechnique Montréal, Montréal, Québec, Canada

[b] CHU Sainte-Justine, Montréal, Québec, Canada.

5.1 Abstract

Purpose. The classification of three-dimensional (3D) spinal deformities remains an open question in adolescent idiopathic scoliosis. Recent studies have investigated pattern classification based on explicit clinical parameters. An emerging trend however seeks to simplify complex spine geometries and capture the predominant modes of variability of the deformation. The objective of this study is to perform a three-dimensional (3D) characterization and classification of the thoracic and thoracolumbar/lumbar scoliotic spine (cross-sectional study). The presence of subgroups within all Lenke types was investigated by analyzing a simplified representation of the geometric 3D reconstruction of a patient's spine, and we propose a new classification approach based on a new machine learning algorithm.

Methods. 3D reconstructions of coronal and sagittal standing radiographs of 663 patients, for a total of 915 visits, covering all types of deformities in adolescent idiopathic scoliosis (single, double and triple curves) and reviewed by the 3D Classification Committee of the Scoliosis Research Society, were analyzed using a machine learning algorithm based on stacked auto-encoders. The codes produced for each 3D reconstruction would be then grouped together using an unsupervised clustering method. For each identified cluster, Cobb angle and orientation of the plane of maximum curvature in the thoracic and lumbar curves, axial rotation of the apical vertebrae, kyphosis (T4–T12), lordosis (L1-S1) and pelvic incidence were obtained. No assumptions were made regarding grouping tendencies in the data nor were the number of clusters predefined.

Results. Eleven groups were revealed from the 915 cases, wherein the location of the main curve, kyphosis and lordosis were the three major discriminating factors with slight overlap between groups. Two main groups emerge among the eleven different clusters of patients: a first with low thoracic deformities and high lumbar deformities, while the other with high thoracic deformities and low lumbar curvature. The main factor that allowed identifying eleven distinct subgroups within the surgical cases (major curves) from Lenke type-1 to type-6 curves, was the location of the apical vertebra as identified by the planes of maximum curvature obtained in both thoracic and thoraco/lumbar segments. Both hypokyphotic and hyperkypothic clusters were primarily composed of Lenke type-1 to type-4 cases, while a hyperlordotic cluster was composed of Lenke type-5 and type-6 cases.

Conclusion. The stacked auto-encoder analysis technique helped to simplify the complex nature of 3D spine models, while preserving the intrinsic properties that are typically measured with explicit parameters derived from the 3D reconstruction.

Keywords: cluster analysis, adolescent idiopathic scoliosis, classification, machine learning,

auto-encoders.

5.2 Introduction

Adolescent idiopathic scoliosis (AIS) refers to a complex deformation of the spine in three-dimensional (3D) space with unknown aetiopathogenesis. Standardized comparisons between treatment strategies or long-term management plans involve a classification system of spinal deformities in order to establish the optimal surgical strategy for example. Ponseti and Friedman made a first endeavor by categorizing spinal curves according to the location and visual patterns of the curve. King et al. proposed to consider the configuration (as observed in the coronal plane), magnitude and degree of flexibility of the scoliosis deformity. Five different curve types were described for spinal arthrodesis recommendations. Their classification system excludes the lumbar segment and the sagittal profile and yields poor validity, reliability and reproducibility. Currently, AIS characterization and treatment recommendations rely mostly on the more comprehensive Lenke classification system (Lenke et al., 2001). A specific curve type, a lumbar spine modifier and a sagittal thoracic modifier define distinctive spine curves. Nevertheless, Lenke classification is based on the conventional measurement of 2D geometric indices such as Cobb angle or central sacral vertebral line. Describing spine deformities with only 2D parameters is insufficient to capture the intricate 3D variability of scoliosis (Labelle et al., 2011).

Classification systems need to improve upon the 2D assessment of scoliosis, which is tied to radiographs in the coronal and sagittal planes. Similar 2D profiles on both coronal and sagittal planes may actually come from different 3D spine geometries (Labelle et al., 2011). A better understanding and characterization of deformation mechanisms should lead to more appropriate treatments and accurate evaluations. The Scoliosis Research Society agreed on a rationalized 3D terminology to describe spinal deformity (Stokes, 1994) and a task force was instructed to assess the clinical relevance and impact of 3D analysis for AIS. Recent efforts have been made to use 3D reconstructions of scoliotic deformities in order to propose accurate and reproducible classification systems, which take into account the 3D nature of the deformity. Numerical methods create new alternatives to current classification systems. First, advanced 3D indices of scoliosis were investigated to discriminate between different types of deformations. Poncet et al. introduced a 3D classification method of scoliotic deformities, based on the geometric torsion of the vertebral body line categorized several curve patterns. Kadoury et al. extended the local geometric torsion measure to regional curves with a parametric curve fitting that was less prone to inaccuracies in the 3D reconstruction. A fuzzy c-means classifier further created subgroups based on the regional geometric

torsion indices. Secondly, regional measures were also explored to provide discriminant indices. Sangole et al. included the axial rotation of the apical vertebrae and the orientation of the plane of maximum curvature (PMC) in the main thoracic (MT) region. Thoracic curve types (Lenke 1) were further subdivided in three different groups with the ISOData algorithm. Duong et al. also considered the orientation of the best-fit plane in the set of 3D parameters. Two different subgroups were found in their small dataset of Lenke 1 curve types. Overall, these studies (Kadoury et al., 2014; Sangole et al., 2009; Duong et al., 2009) share a similar framework. Their classification systems are derived from the clustering of hand-engineered parameters, which were calculated from 3D spine reconstructions. However, relying on geometric indices sets out on a quest in search of the best characteristics to describe the 3D nature of scoliotic spines.

Numerical methods should be able to capture within a simplified space, the highly dimensional and complex nature of a fully geometric 3D reconstruction of the spine, both on a regional (spinal) and local (vertebra) levels. This implies directly analyzing the 3D spine models instead of expert-based features as it has been experimented previously. Duong et al. proposed a wavelet-based compression technique of the spinal curves. Kadoury and Labelle investigated a manifold learning algorithm based on locally linear embedding for dimensionality reduction of 3D spine models of the Lenke 1 curve types. However, these local techniques for dimensionality reduction tend to suffer from the curse of dimensionality and to be sensitive to data models which tend further away from the general trend of the normal distribution (van der Maaten et al., 2009). Hence, increasing the number of landmarks to describe the 3D spine models or including other Lenke types will lead to miss-classification of an important number of samples. Global nonlinear techniques for dimensionality reduction could overcome these drawbacks (van der Maaten et al., 2009) by preserving the global properties of the 3D spine models.

In this study, we propose to use recent advances in artificial intelligence to simplify the high-dimensional and complex nature of geometric 3D spine reconstructions for classification purposes. This highly non-linear transformation discriminates between AIS scoliotic curves by learning the intrinsic properties of 3D spine reconstructions by preserving the global properties. Once a low-dimensional representation has been learned from a cohort of 3D spine models, new classes can be derived from their simplified description.

5.3 Materials and methods

We evaluate the relevance of machine learning algorithms, namely stacked auto-encoders, on a large database that comprises 915 reconstructions of all Lenke types (i.e. from Lenke 1 to

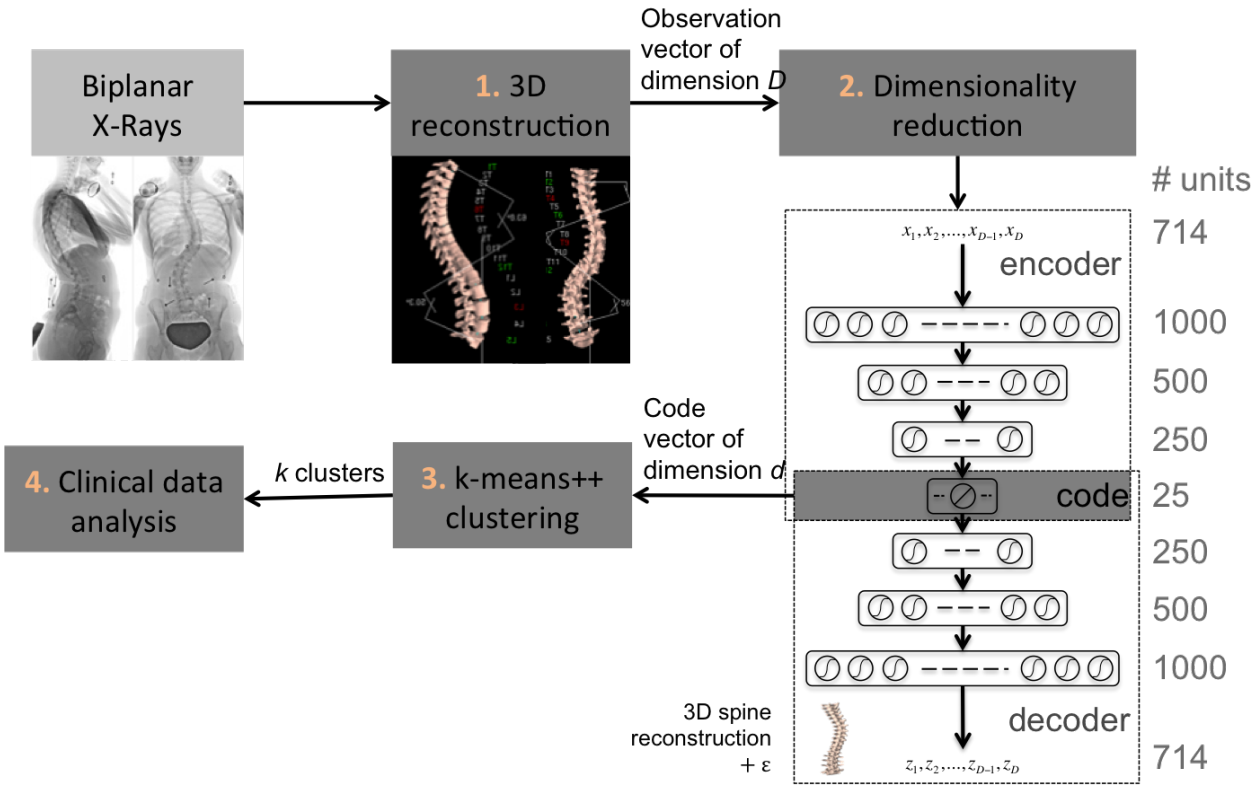


Figure 5.1 Flowchart of the classification method. The system sequentially: (1) reconstructs a 3D spine model, x of size D, from biplanar X-rays for each patient; (2) maps the high-dimensional spine reconstruction to a low-dimensional space, called a code, with stacked auto-encoders of symmetric layer sizes which continuously compresses the code to a smaller dimension d; (3) clusters the low-dimensional spines into k sub-groups; (4) validates the cluster relevance with the clinical data.

Lenke 6). The proposed framework is illustrated in Figure 5.1 and consists of four main steps: (1) reconstruction of a 3D spine model from biplanar X-rays for each patient; (2) encoding each 3D spine in a low-dimensional space; (3) clustering of the encoded spines models; (4) validation of the sub-groups with clinical data.

5.3.1 Patient Data

In this retrospective cohort study, data of 663 preoperative AIS patients from nine scoliosis centers (New York City, Baltimore, Philadelphia, CHOP, Miami, San Diego, Wilmington, Montreal, Vancouver) during an 18-year period (1994-2012) were selected. From this group, 151 patients had repeat measurements from multiple clinic visits (mean = 2.7 visits), yielding a total of 915 cases. All patients were diagnosed with an adolescent idiopathic scoliosis in the thoracic and/or lumbar spine. The mean of the major Cobb angle was 58.8±15.2° (range

= 21.3–113.6°). Note that a major Cobb angle corresponds to the maximum value between the main thoracic (MT) Cobb angle and the thoracolumbar/lumbar (TLL) Cobb angle, both measured in the plane of maximum curvature (PMC). Members of the 3D Classification of the SRS assigned a Lenke type to all cases, which are divided in: 312 Lenke type-1, 118 Lenke type-2, 152 Lenke type-3, 122 Lenke type-4, 113 Lenke type-5, and 98 Lenke type-6 curves.

5.3.2 Three-Dimensional Reconstruction of the Spine

The spine was reconstructed in 3D from calibrated coronal and sagittal radiographs of the patient in a standing position (Kadoury et al., 2009) . Radiographs were acquired from either the EOS low dose imaging device (EOS imaging, Paris, France) or from a conventional radiographic imaging system. A statistical model from a database of scoliotic patients was used to reconstruct an initial spine model in 3D. Anatomical landmarks on each vertebra were further refined with an iterative process based on several features extracted from the radiographs. Finally, an experienced user at our institution corrected and validated the anatomical landmark positions to generate a personalized 3D spine for each patient. A 3D reconstruction of the spine consists of fourteen anatomical landmarks per vertebra (12 thoracic, 5 lumbar): center, left, right, anterior and posterior of both superior and inferior vertebral endplates (10 landmarks); and tips of both pedicles (4 landmarks). All 3D spine models were normalized with regards to their height and rigidly translated to a common referential at the L5 vertebra. Hence, each model contains 714 features, which corresponds to the concatenation of the 14 landmarks with 3 dimensional coordinates (x, y, z), identified on each of the 17 vertebrae.

5.3.3 Encoding of Three-Dimensional Spine Models

The geometric 3D spine models were then simplified into a low-dimensional encoding in order to capture the main factors of variation in the shape of the spine from the given cohort. To perform this step, a stacked auto-encoder (SAE) was used to simplify the representation of the 3D reconstructions. A SAE consists of a specific artificial neural network architecture that learns a latent representation of the inputs. The algorithm is performed as follows. Each spine is represented as an input high-dimensional vector of 3D coordinates for all the landmarks of a spine model (denoted as x), and the SAE first attempts to compress each spine into a latent representation of low dimension using an encoder function. The representation in low-dimension is considered as a compressed version of the input, called a code (denoted as c). Once a code is obtained, the algorithm will decode this compressed vector and regenerate

an output 3D spine reconstruction (denoted as y). This encoding procedure is optimized by minimizing in an iterative fashion the difference between the inputs x and outputs y. Stacking several auto-encoders helps the artificial neural network to become invariant to most local changes and disentangle the main factors of variation in the dataset (Vincent et al., 2010, 2008).

5.3.4 Clustering

Once a large database of 3D reconstructed spine models were encoded into low-dimensional codes, the k-means++ clustering algorithm (Arthur and Vassilvitskii, 2007) partitioned the spine dataset into k separate sub-groups. This clustering algorithm is a variant of the traditional k-means clustering algorithm that integrates a probabilistic seeding initialization method. The selection of the right number of clusters k is based on the validity ratio (Ray and Turi, 1999) , which minimizes the intra-cluster distance and maximizes the inter-cluster distance.

5.3.5 Statistical Analysis

We validated the clustered data points with standard geometrical indices in the main thoracic (MT) and thoracolumbar/lumbar (TLL) regions. For each spine, the Cobb angles and the orientations of the PMC were computed in both regional curves. The kyphotic angle was measured between T2 and T12 on the sagittal plane. The lumbar lordosis angle was defined between L1 and S1 on the sagittal plane. The axial rotation of the apical vertebra in the MT region was computed by the Stokes method Stokes et al. (1986) . Finally, the pelvic incidence (PI) was measured between the line perpendicular to the sacral plate at its midpoint and the line connecting this point to the axis of the femoral heads (Legaye et al., 1998). One-way ANOVA tested differences between the cluster groups with a significance level $\alpha = 0.05$. The p-values were adjusted with the Bonferroni correction. Moreover, an experienced surgeon at our institution performed a clinical assessment of the ten closest cases near the centroid of each cluster.

5.4 Results

The cohort of 915 visits from 663 patients was randomly divided into a training set (645 cases), a validation set (135 cases) and a testing set (135 cases) for unbiased evaluation. In order to determine the hyper-parameters of the neural network, an exhaustive grid search was performed on the validation set by minimizing the mean squared error. The architecture

yielding the lowest error is presented in Figure 5.1. We used an encoder with four latent layers of size (layer 1: 1000 units; layer 2: 500 units; layer 3: 250 units; code: 25 units) and a decoder with tied weights to map the high-dimensional patient's spine models into low-dimensional codes. Weight parameters were initialized by a denoising auto-encoder to capture the statistical dependencies between the inputs. The final model was trained by using the entire dataset of 915 cases. Note that 5 cases were further excluded because the pelvis radiograph was not available.

The k-mean++ clustering detected eleven different groups from the low-dimensional encoding of 3D geometrical models based on the validity ratio. Table 5.1 presents the clinical statistical data analysis for these eleven groups. The mean values of all geometric parameters are listed for all eleven groups and the differences between all groups were found to be statistically significant ($\alpha = 0.05$) for each parameter. Table 5.2 presents the Lenke curve type distribution across the eleven clusters, while Table 5.3 offers a summary description of each cluster based on the observed parameters.

Figure 5.2 presents sample cases for all these eleven clusters detected by the classification framework. In order to visualize the distribution of samples in this low-dimensional space, a principal component analysis was performed on the encoded samples of 25 dimensions, in order to project the encoded spine reconstructions to 3D and 2D views. Figure 5.3 depicts the visualization of the first three principal components (PC) from this analysis. The first PC explains 46% of the variance in the encoded geometric spines, representing the location of the major curve. High values in the first PC tend to increase angulation of TLL plane and the axial rotation angle of the apical vertebra while decreasing the Cobb angle and the angulation of MT plane. The second PC explains 26% of the variance and is related to the lordotic angle. The third PC explains 11% of the variance and is related to the kyphotic angle.

5.5 Discussion

In this 3D analysis of spinal deformities, a novel method simplifying the representation of the geometric 3D reconstruction of a patient's spine was presented to develop a new classification approach based on a novel machine learning algorithm. Previous systems based on 2D radiographic images covered all types of curve patterns and provided a reliable set of measures which take under account the deformity in the sagittal plane, along with specific modifiers (King et al., 1983; Lenke et al., 2001). Still, relying on 2D projections of a complex 3D curve as encountered in AIS represents a considerable limitation to these standard approaches. On the other hand, evaluating the deformity based on discrete local 3D measure-

Table 5.1 Mean and standard deviation values of the geometric parameters in the MT and TLL regions, within all eleven clusters detected by the proposed classification framework. Red represents the maximum value for all identified clusters; green represents the minimum value for all identified clusters. MT: Main thoracic, TLL: Thoracolumbar/lumbar, PMC: plane of maximal curvature

	Parameters (all in degrees)							
Cluster	MT Cobb	MT Rot.	TLL Cobb	TLL Rot.	Kypho-sis	Lordosis	Axial Rot.	Pelvic inc.
I (n=114)	58±11	80±8	43±12	56±15	20±11	-60±9	-22±9	57±10
II (n=118)	56±13	72±10	43±13	58±18	38±12	-67±10	-19±9	52±10
III (n=77)	39±16	58±16	54±15	81±17	40±11	-69±10	-5±12	51±9
IV (n=33)	47±12	74±17	54±14	74±19	17±14	-65±10	-11±8	68±14
V (n=106)	53±9	77±8	38±11	49±16	29±10	-62±13	-20±10	51±11
VI (n=93)	77±13	83±8	53±15	61±15	19±19	-57±14	-29±10	54±10
VII (n=111)	42±13	70±13	50±13	77±14	26±10	-61±11	-8±9	54±12
VIII (n=55)	72±11	71±8	57±14	73±15	40±11	-65±10	-25±11	51±9
IX (n=68)	45±11	80±10	39±13	67±18	20±9	-57±10	-13±9	55±9
X (n=62)	68±12	81±9	40±11	44±14	33±12	-62±13	-25±9	50±8
XI (n=73)	38±14	62±17	52±14	93±14	35±11	-61±16	-6±10	48±11
p-value	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

ments, such as axial rotation or geometric torsion (Kadoury et al., 2014), is inevitably linked to the quality of the 3D reconstruction and to the inter-rater variability of these pre-defined measurements. In this paper, we attempt to achieve a classification of 3D patterns based on the global representation of the spine without using explicit parameters derived from the 3D reconstruction of the spinal shape. The approach was able to detect eleven sub-groups based on their low-dimensional representation. The differences in clinical measurements (Cobb angles and orientation of PMC, kyphosis, lordosis, pelvic incidence) between all these new 3D sub-groups were found to be statistically significant.

Two clinically relevant groups emerge among the eleven different clusters of patients detected by the algorithm. In the first, clusters VII, XI, and III (illustrated as shades of blue in Figure 5.2) represent the clusters with low thoracic deformities and high lumbar deformities. An increase of the TLL Cobb angle in the PMC and a decrease of the axial rotation in the

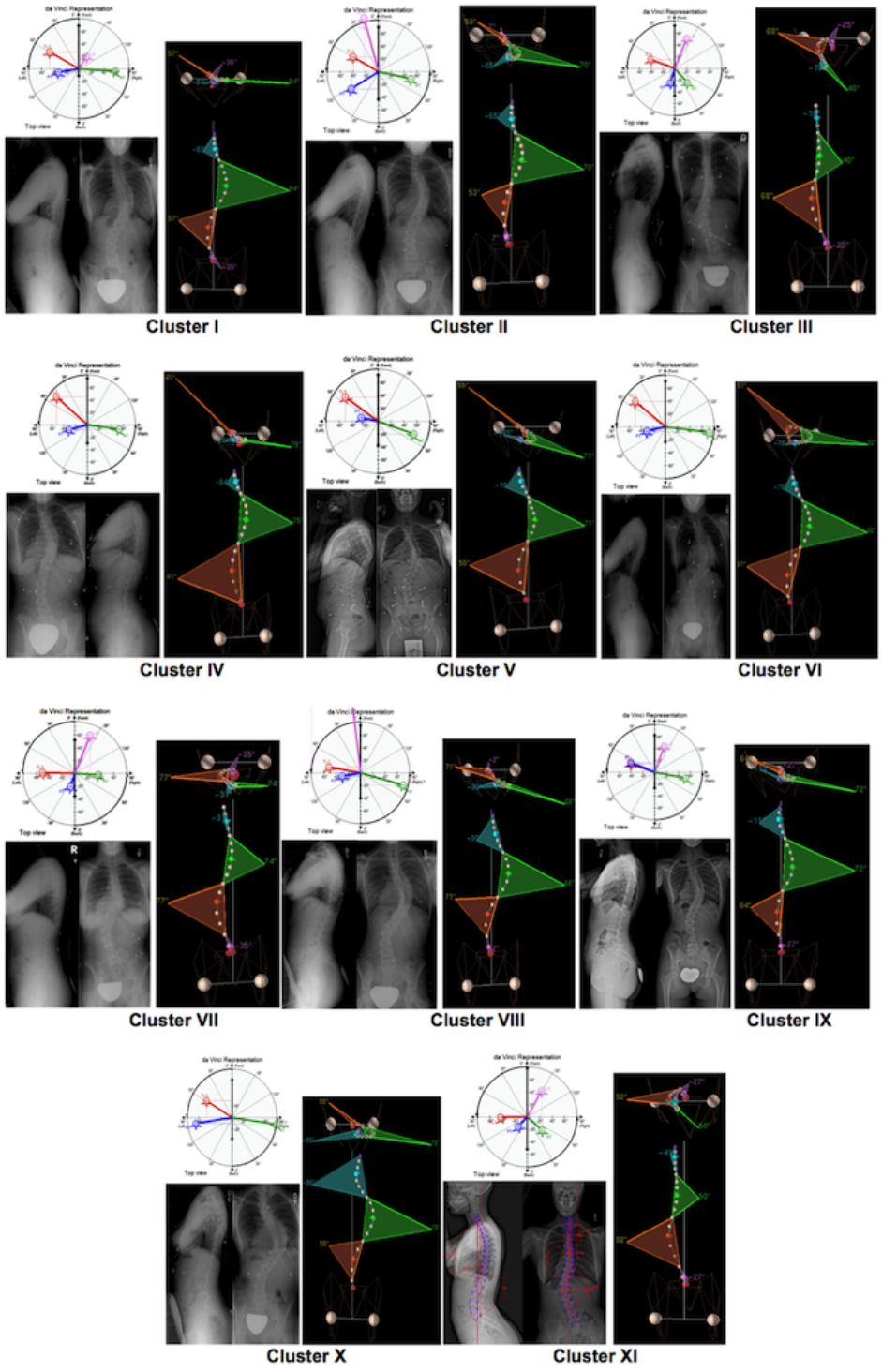


Figure 5.2 Sample cases for each of the eleven clusters found by the clustering algorithm. For each cluster sample, coronal/sagittal radiographs, daVinci representations (Labelle et al., 2011), coronal and top views of the 3D reconstruction model are presented.

Table 5.2 Composition of Lenke sub-types in percentages (%) for each detected cluster.

Cluster	Lenke I	Lenke II	Lenke III	Lenke IV	Lenke V	Lenke VI
I (n=114)	52.6%	18.4%	20.2%	8.8%	0.0%	0.0%
II (n=118)	33.1%	21.2%	26.3%	16.1%	1.7%	1.7%
III (n=77)	7.8%	2.6%	6.5%	10.4%	44.2%	28.6%
IV (n=33)	6.1%	3.0%	33.3%	3.0%	24.2%	30.3%
V (n=106)	67.9%	13.2%	8.5%	9.4%	0.0%	0.9%
VI (n=93)	46.2%	12.9%	17.2%	23.7%	0.0%	0.0%
VII (n=111)	14.4%	4.5%	11.7%	9.0%	31.5%	28.8%
VIII (n=55)	30.9%	9.1%	23.6%	29.1%	0.0%	7.3%
IX (n=68)	38.2%	8.8%	25.0%	13.2%	4.4%	10.3%
X (n=62)	37.1%	41.9%	6.5%	12.9%	1.6%	0.0%
XI (n=73)	9.6%	0.0%	12.3%	11.0%	39.7%	27.4%

MT region from cluster VII to cluster III are clearly apparent from Table 5.1. On the other hand, the MT Cobb angles in the PMC remain low. A high distribution a Lenke 5 and 6 curve types in these three clusters confirm these patterns. In the second group, clusters II, V, I and X (illustrated in shades of red/orange in Figure 5.2) represent the clusters with high thoracic deformities and low lumbar deformities. An increase of the MT Cobb angle and the MT orientation in the PMC, from cluster II to cluster X, is observable. A similar behavior for the axial rotation of the apical vertebra in the MT region is observed. The geometrical parameters obtained in the lumbar segment remain low. An absence – or a very small presence – of Lenke 5 and 6 curve types in these five clusters confirms these patterns. These observations reveal the fact that the location of the major curve (thoracic, lumbar, thoraco-lumbar/lumbar) is the most discriminant clinical factor in distinguishing different classes of deformity. Within these two groups, there exists an important range of kyphotic and lordotic profiles, as well as a spectrum of varying curve severity that is observable, thus suggesting that there exists variability with single or double major curves, either in the thoracic and lumbar regions.

Clusters can also be stratified based on their kyphotic and lordotic profiles. Clusters II, III and VIII represent the clusters with hyper-kyphotic and hyper-lordotic profiles. However,

Table 5.3 Cluster descriptions for the eleven clusters detected by the stacked auto-encoder framework.

Cluster	Cluster description
I	High MT PMC orientation, Hypokyphotic, Hypolordotic, High PI
II	Medium-High MT PMC orientation, Hyperkyphotic, Hyperlordotic
III	High TLL PMC Cobb, High TLL PMC orientation, Hyperkyphotic, Hyperlordotic
IV	High TLL PMC Cobb, Hypokyphotic, Hyperlordotic, High PI
V	High MT PMC orientation, Low TLL PMC orientation
VI	High MT PMC Cobb, High MT PMC orientation, High TLL PMC Cobb, Hypokyphotic, Hypolordotic, High axial rotation of apical vertebra
VII	High TLL PMC Cobb, High TLL PMC orientation
VIII	High MT PMC Cobb, High TLL PMC Cobb, Hyperkyphotic, Hyperlordotic, High axial rotation of apical vertebra, Low PI
IX	High MT PMC orientation, Hypokyphotic, Hyperlordotic
X	High MT PMC Cobb, High MT PMC orientation, Hyperlordotic, High axial rotation of apical vertebra
XI	High TLL PMC Cobb, Very High TLL PMC orientation, Low PI

cluster II and cluster III have completely different deformities in their respective MT and TLL regions. Cluster VIII denotes the cluster with high deformities in both MT and TLL regions. This is confirmed by the highest percentage of Lenke type-4 from all clusters (29%). Clusters I, VI and IX represent the clusters with hypo-kyphotic and hypo-lordosis profiles. Cluster VI has the highest thoracic deformities and relatively high lumbar deformities. This behavior is similar to the cases included in cluster VIII. Cluster IX differs from cluster VI with lower Cobb angles in both MT and TLL regions in the PMC. Finally, cluster IV represents the cluster with hypo-kyphotic and hyper-lordotic profiles with high thoracic deformities. These findings confirm also the existence of hypo-kyphotic profiles (clusters IV and VI) within groups exhibiting high thoracic deformities. The difference between these two clusters is the major difference in angulation and Cobb angle of the plane of maximal deformity in the thoracic region, thereby suggesting that regional angulation is still an important factor in assessing the deformation.

We used an automated classification method using a deep learning technique, namely with stacked auto-encoders, to discover sub-groups within a large pool of patients with both thoracic and lumbar deformations. The code layer of the auto-encoder learns a distributed representation in low dimension that aims to capture the main factors of variation in the clinical dataset. However, different examples from the distribution of the training dataset may potentially yield to high reconstruction errors. Therefore, having a large and representative training dataset of AIS is critical. This will also prevent the model from overfitting. The current study evaluated the 3D sub-classification of all Lenke types for thoracic and lumbar

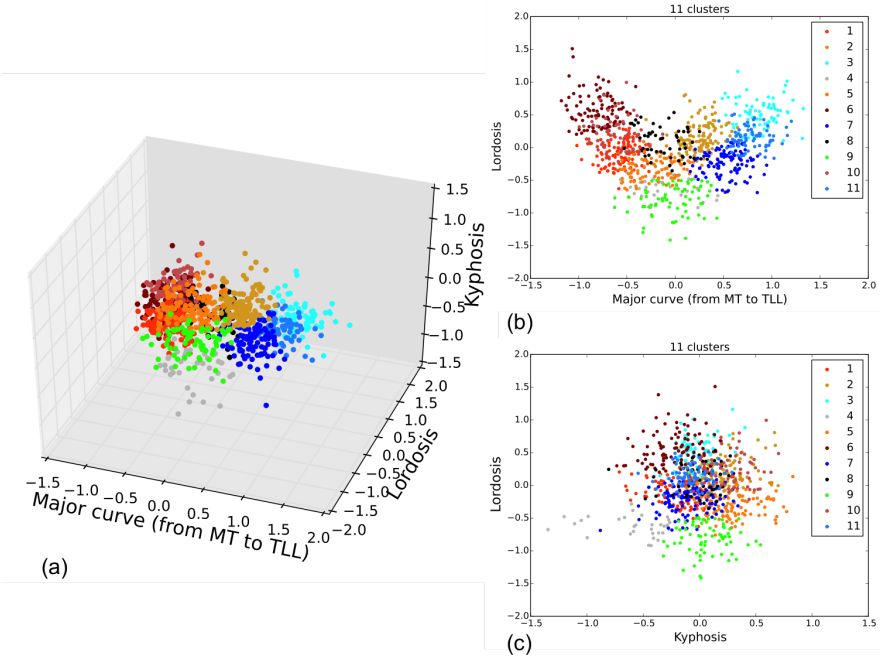


Figure 5.3 Visualization of the eleven clusters found by the k-means++ algorithm from the low-dimensional encoding of 3D geometrical models. Each color point represents a single 3D spine reconstruction in a low-dimensional space. (a) 3D scatter plot of all 915 cases in the low-dimensional space using principal component analysis. The 3D view is projected onto 2D views with (b) First and second principal components, and (c) Second and third principal components.

scoliotic curves, suggesting that shape variability is present within an existing 2D group used in clinical practice. However, these types of approaches include complex synthetization tasks, which require sizeable datasets to improve the data representation within the code layer. Therefore, a larger multicentric dataset may help to significantly increase the number of cases from various sites and obtain a more reproducible model. Furthermore, the development of computational methods will ultimately lead to more reliable classification paradigms, helping to identify possible cases which might progress with time. Future work will use longitudinal data for surgical treatment planning, whereas each case is considered independently in the current framework. Finally, a reliability study will be undertaken to evaluate the clinical relevance of the classification system in terms of surgical strategy.

5.6 Acknowledgements

This paper was supported in part by the CHU Sainte-Justine Academic Research Chair in Spinal Deformities, the Canada Research Chair in Medical Imaging and Assisted Interven-

tions, the 3D committee of the Scoliosis Research Society, the Natural Sciences and Engineering Research Council of Canada and the MEDITIS training program. The authors would like to thank the developers of Theano (Bergstra et al., 2010; Bastien et al., 2012).

CHAPITRE 6 ARTICLE #3 : CONVOLUTIONAL NETWORKS FOR KIDNEY SEGMENTATION IN CONTRAST-ENHANCED CT SCANS

Cet article présente l'utilisation de réseaux de neurones pour l'apprentissage supervisé de représentations d'images médicales. Un réseau à convolution apprend à distinguer le rein des autres structures anatomiques présentes dans les images tomodensitométriques abdominales d'une large banque de données de données (63 patients pour 79 images tomodensitométriques).

Auteurs

William Thong[a], Samuel Kadoury[a], Nicolas Piché[b], Christopher J. Pal[a]

Affiliations

[a] Polytechnique Montréal, Montréal, Québec, Canada

[b] Object Research System Inc. (ORS), Montréal, Québec, Canada

6.1 Abstract

Organ segmentation in medical imaging can be used to guide patient diagnosis, treatment and follow-ups. In this paper, we present a fully automatic framework for kidney segmentation with convolutional networks (ConvNets) in contrast-enhanced CT scans. In our approach a ConvNet is trained using a patch-wise approach to predict the class membership of the central voxel in 2D patches. The segmentation of the kidneys is then produced by densely running the ConvNet over each slice of a CT scan. Efficient predictions can be achieved by transforming fully-connected layers into convolutional operations and by fragmenting the maxpooling layers to segment a whole CT scan volume in a few seconds. We report the segmentation performance of our framework on a highly-variable dataset of 79 cases using a variety of evaluation metrics.

6.2 Introduction

Organ segmentation in medical images plays an important role in clinical diagnosis, radiotherapy planning, interventional guidance and patient follow-ups. In recent years there have been a strong development of computer-assisted tools to help clinicians in this tedious and time-consuming task. However, organ segmentation copes with challenges emerging from inherent hardware acquisition noise, imaging artifacts, patients and vendor variability, etc. Consequently, segmentation algorithms must be robust and accurate enough to generalize to unseen medical images.

Several methods for kidney segmentation have been investigated in computerized tomography (CT) scans in the literature. *Registration-based approaches* are very common and very popular in the medical imaging, especially for brain imaging. The segmentations are obtained by warping an atlas to the unlabelled target image. This simple procedure heavily relies on the quality of the atlas, which usually doesn't achieve good performance in inter-subject registration (Criminisi et al., 2013). Wolz et al. proposed a hierarchical multi-atlas technique to overcome this issue. A global-to-local scheme produced the segmentation of several abdominal organs. Several atlases were weighted according to image appearance at a global and organ level for the registration to the target image. A patch-based method finally refined the segmentation. Chu et al. extended this global-to-local approach by diving the image volume space into many sub-spaces. The final organ segmentation was obtained with graph-cuts. However, the major drawbrack of these registration-based approaches comes from the multiple registrations that are time-consuming. Several hours are usually required to generate the segmentations (Wolz et al., 2012; Chu et al., 2013). *Random fields* can incor-

porate structure in the models by adding voxel neighborhood information. Freiman et al. used a graph min-cut approach to automatically segment the kidneys by globally optimizing the edge weights in the graph. Khalifa et al. also used random fields as a prior for their level set approach. However, these iterative methods based on random fields also suffer from expensive computation time. With the increasing availability of medical images, *supervised learning approaches* have gained a recent interest for organ detection (Criminisi et al., 2013) and segmentation. Cuingnet et al. proposed a three-step procedure. Regression forests predicted bounding box coordinates of both kidneys. Classification forests then assigned a probability to each voxel inside the bounding boxes. A template of an ellipsoidal shape was finally deformed to create the segmentations. Glocker et al. extended this idea of combined regression and classification for multi-organ segmentation in the abdomen. They derived a new objective function to train random forests that incorporated both class membership and spatial position. However, both random fields and random forests approaches greatly depend on the selected input features.

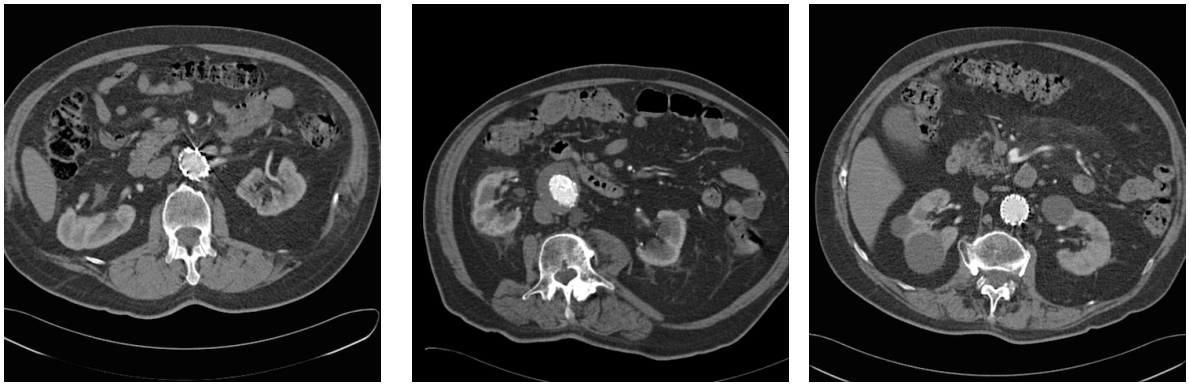
Since their dramatic success in the ImageNet challenge (Krizhevsky et al., 2012), there has been a resurgence of interest in the use of convolutional networks (ConvNets) for a variety of computer vision tasks. Learning multiple levels of representation better captures the rich and complex variability in the data than hand-engineered features (LeCun et al., 1998). Hence, ConvNets have tremendous potential to overcome some of the major challenges in biomedical image segmentation. In this paper, we present a unified and automatic framework based on a ConvNet for kidney segmentation in contrast-enhanced CT scans. Our contribution in this work is to demonstrate that a ConvNet approach for organ segmentation in medical imagery is able to produce high quality, pixel level results in about a small amount of time.

6.3 Methods

We define the kidney segmentation problem as a classification-based task. Given a 2D slice $\mathcal{I}$ of a CT scan volume $\mathcal{V}$, the objective is to predict the class membership — *background* or *kidney* — for each pixel $i \in \mathcal{I}$. A framework based on convolutional networks was designed to learn this objective. Given a training dataset of segmented kidney images (Sec 6.3.1), a ConvNet was trained patch-wise (Sec. 6.3.2). Unique square patches p, centered around i, were randomly sampled in each CT scans. After training, the ConvNet architecture was transformed by simple modifications (Section 6.3.3), without altering the learned weights, to produce the prediction at the pixel-level for every slice $\mathcal{I} \in \mathcal{V}$. Hence, the modified ConvNets took as input the entire slice $\mathcal{I}$, which is more efficient than sliding patches of fixed-size over $\mathcal{I}$. Two different transformations were used. The fully-connected layers were

turned into convolutional layers in *ConvNet-Coarse* to produce a coarse segmentation at a smaller resolution. Furthermore, the maxpooling layers were fragmented in *ConvNet-Fine* to generate a fine segmentation at the same resolution of the input slice at the cost of a higher computation time.

6.3.1 Dataset and pre-processing steps



(a) (b) (c)

Figure 6.1 Sample slices showing the variability of the dataset: (a) metallic artifact from the endovascular stent; (b) displacement of several organs; (c) apparent cyst in both kidneys.

We collected 79 contrast-enhanced CT scans from different research hospitals, different CT scanner manufacturers and different acquisition parameters. The dataset came from 63 patients and was completely anonymized. Most of the patients received an endovascular stent-grafting for the treatment of abdominal aortic aneurysms, which produced strong metal streak artifacts in the images. Both kidneys in each scan were manually segmented by an expert user using a commercial software application. Figure 6.1 depicts the high variability of the CT scans. The axial resolution ranges from 0.58 to 0.92 mm^2 and the slice thickness ranges from 2 to 3 mm. The imaging matrix comprises 512×512 pixels in the axial plane and 100 to 264 slices. The images were windowed between -150 and 250 Hounsfield units to isolate the volume body from the regions composed of air and exclude bone regions. The dataset was further standardized to have zero-mean and unit-variance slice-wise.

6.3.2 Training ConvNet

Figure 6.2 illustrates the ConvNet architecture used for the training procedure. It consists of two convolutional layers, with a kernel size of 4 and 3 respectively, followed by two fully-

connected layers, both composed of 256 hidden units. Each convolutional layer was followed by a non-overlapping maxpooling layer, with a kernel size of 2.

Fifty thousand unique patches of size 43×43 were randomly sampled in the axial plane of each case with balanced classes, i.e. 50% of the patches belonged to the *background* and 50% to the *kidney*. This yielded a total of 1.95 million training examples. Each patch was labelled according to the class membership of the central voxel. For each weights update, patches were randomly rotated (by an angle between $[-20, 20]^\circ$) and randomly flip horizontally (with a probability of 0.5). Weights were updated using Nesterov accelerated gradient with an initial momentum of 0.9 and an initial learning rate of 0.03. All layers used rectified linear units and were initialized following Glorot and Bengio (2010). Dropout with a probability of 0.5 was applied to all fully-connected layers as a regularizer.

6.3.3 Segmentation with ConvNets

A typical image volume $\mathcal{V}$ contains dozens of millions of voxels. For example, given an image volume of size 512x512x200, the ConvNet would have to predict the class membership of more than 52 million of patches. In other words, this exhaustive sliding-window search scheme would be too computationally expensive for computer-aided diagnosis systems. In order to effectively address this matter, the fully-connected layers were turned into convolutional layers (Sermanet et al., 2013) in *ConvNet-Coarse* and maxpooling layers were fragmented (Giusti et al., 2013) to densely run *ConvNet-Fine* at each location in the input space.

ConvNet-Coarse.

Convolutional layers incorporate by nature a sliding-window scheme: contiguous units in a feature map share the same set of weights (i.e. the same computation) resulting in overlapping receptive fields (LeCun et al., 1998). This inherent property can be exploited to process variable input size by transforming the fully-connected layers into convolutional layers (Sermanet et al., 2013). Thus, the fully-connected layers of the ConvNet described in Sec. 6.3.2 will be transformed into two convolutional layers with a kernel of size 9 and 1 respectively. Instead of producing a binary output corresponding to the class membership, the ConvNet now outputs a spatial probability map in a single forward propagation where each element corresponds to a specific field-of-view (FOV) of the input image. However, each of these FOVs are separated by a stride corresponding to the overall subsampling factor that comes from the maxpooling layers. Consequently, the outputs of *ConvNet-Coarse* have the quarter of the size of the input due to the two non-overlapping maxpooling layers with a kernel size

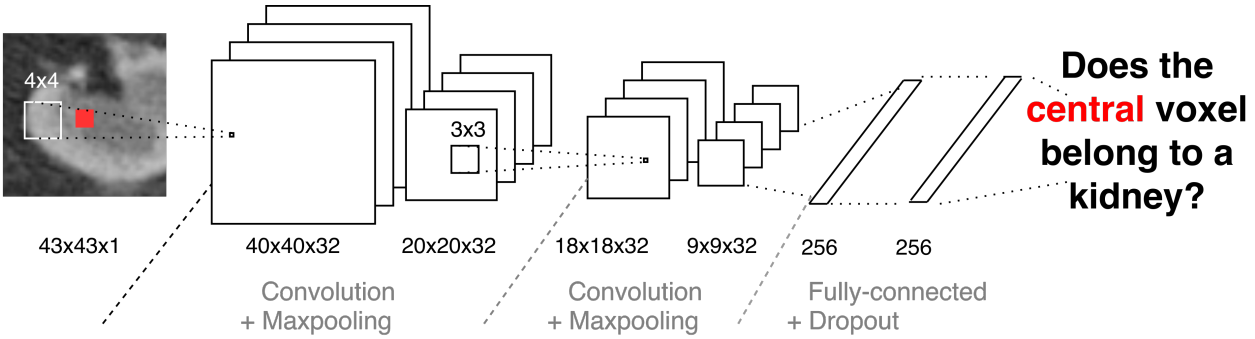


Figure 6.2 ConvNet architecture used during the training stage.

of 2. *ConvNet-Coarse* then produces an output of 128×128 [1] for an input of size 512×512. Therefore, the spatial probability maps of *ConvNet-Coarse* need to be upscaled with a factor 4 to get the kidney segmentation at the original sampling. A nearest neighbor interpolation and a bilinear interpolation were assessed in this framework.

ConvNet-Fine.

A maxpooling layer splits a feature map (i.e. the ouput of a convolutional layer) into sub-regions and outputs the maximum local value for each sub-region. This means that the output of a maxpoolng layer has a lower resolution than its input. Typically, a ConvNet contains multiple convolutional and maxpooling layers, yielding an output that is subsampled several times. A resolution augmentation can be performed to overcome this drawback by applying the maxpooling operation at several offsets to cover all the possible combinations of sub-regions (Giusti et al., 2013). Figure 6.3 illustrates an example of the principle. For a kernel of size 2, the feature map is shifted by an offset of $\{0, 1\}$ in both axis, which corresponds to a set of 2D offsets equivalent to $\mathbf{o} = \{0, 1\} \times \{0, 1\} = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. Therefore, the maxpooling operation is repeated 2^2 times for every 2D offset combinations, producing a total 4 different output fragments. With two non-overlapping maxpooling layers, the model described in Sec. 6.3.2 will generate 16 fragments. Each of these fragments represents a spatial probability map where each element in a fragment corresponds to a different FOV. By interleaving all the fragments, a dense spatial probability map can be reconstructed at the original input resolution.

As a post-hoc processing, an alternating sequential filter with 3D opening and closing operators of size 3 was applied to remove sharp edges and break narrow strips between 3D

1. Note that the output size is actually 118×118 because no padding was used before applying the convolutional operations.

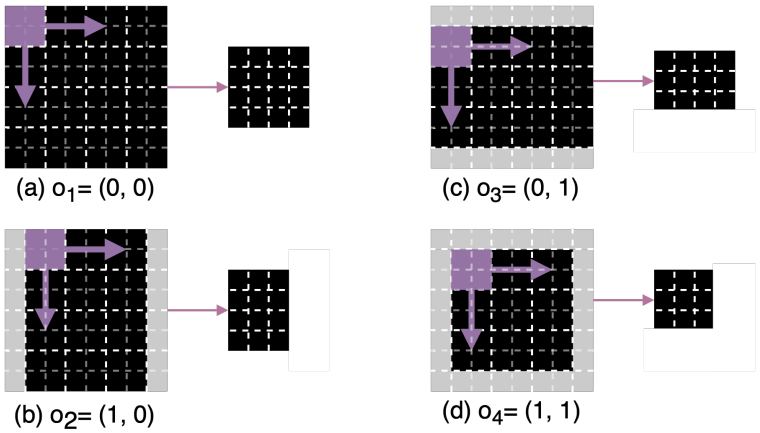


Figure 6.3 2D fragmentation of a 8×8 feature map by a non-overlapping maxpooling operator with a kernel size of 2. The maxpooling operation is shifted in both axes by a set of 2D offsets $\mathbf{o} = \{(0,0),(1,0),(0,1),(1,1)\}$, yielding a total of 4 different fragments. Note that o_1 in (a) is equivalent to a traditional maxpooling layer with no offset.

components. The kidneys then correspond to the two biggest connected components in the 3D volume.

6.4 Results

From the dataset of 79 scans, three random subsets were generated: a training set (39 cases coming from 33 patients), a validation set (20 cases from 15 patients), and a testing set (20 cases from 15 patients). If a patient had multiple visits, all the visits were constrained to be in the same subset for unbiased evaluation. Hyper-parameters of the ConvNet were selected based on the error rate on the validation set. The probability map threshold was chosen according to the validation set. Evaluation metrics are reported on the testing set.

Figure 6.4 depicts the output of both ConvNets (*ConvNet-Coarse* and *ConvNet-Fine*) as well as the final output of the framework. Some slight differences at the edges can be observed. Table 6.1 presents the median, 1st quartile and 3rd quartile of the quantitative metrics values for the segmentation of the 20 cases in the testing set compared to the manual segmentation. The framework missed to detect and segment two kidneys over the 40 kidneys present in the testing set. We evaluated the segmentation performance using the Dice coefficient (DC), the Hausdorff distance (HD), the average symmetric surface distance (ASSD), and the precision and recall scores. Note that we used a more robust version of the HD by taking the 95^{th} percentile instead of the maximum distance surface distance between two sets. Computation times corresponds to the time taken by the ConvNet to predict a 512×512 slice with an

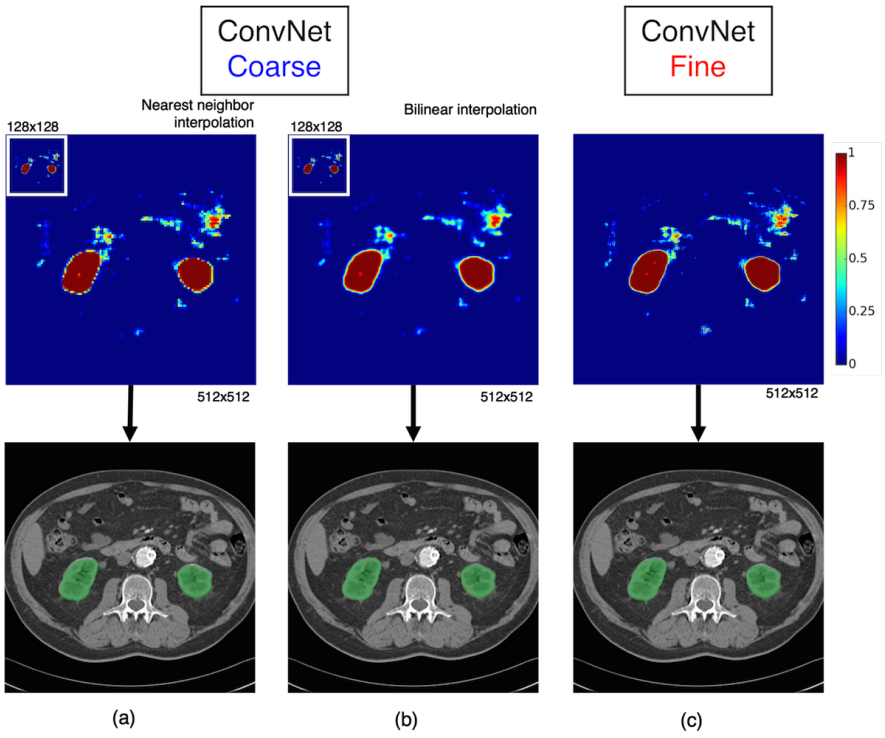


Figure 6.4 Sample slices illustrating the results of the framework for kidney segmentation. The top row shows the output of the ConvNet while the bottom row shows the final segmentation after the post-processing steps. (a) and (b) represent the results coming from *ConvNet-Coarse*. The upper left image at the top row is actually the output at original size that is further rescale by nearest neighbor and bilinear interpolation in (a) and (b) respectively. (c) represents the result of *ConvNet-Fine*.

NVIDIA Tesla C2070.

We observe that our coarse scale predictions are surprisingly accurate when scaled up. The Dice scores and recall for the coarse model are in fact higher than the fine scale model; however the precision is slightly lower. We believe this effect arises from the fact that kidneys are relatively smooth organs and therefore our interpolation technique is able to yield very high quality segmentations.

6.5 Discussion

This paper has presented a fully automatic framework for automatically detecting and segmenting the left and right kidneys in contrast-enhanced CT scans based on convolutional networks. We have used an approach that allows for high quality predictions to be produced at a lower resolutions more rapidly. Due to our coarse-fine design and the prediction qual-

Table 6.1 Segmentation evaluation of the left and right kidneys by *ConvNet-Coarse* with linear interpolation and *ConvNet-Fine* on the testing set of 20 scans. The median, 1st and 3rd quartiles values are reported.

	ConvNet-Coarse		**ConvNet-Fine**	
	left	*right*	*left*	*right*
Dice (%)	94.53 91.72–95.04	93.07 89.99–94.28	93.62 91.99–94.98	92.52 88.83–94.47
ASSD (mm)	0.90 0.65–1.42	1.11 0.87–2.08	1.00 0.66–1.44	1.23 0.85–2.46
HD (mm)	4.97 2.80–8.31	6.35 4.24–17.90	4.84 2.71–9.51	7.02 4.12–17.50
Precision (%)	94.08 89.51–95.11		94.66 89.23–95.43	
Recall (%)	95.09 92.53–96.39		93.78 90.79–95.61	
Time (ms)	35.3 ± 0.3		223.5 ± 0.2	

ity at the coarse resolution, our approach is able to generate predictions at variable pixel densities or resolutions. This can be implemented in a way that allows us to further accelerate prediction speed in regions that do not contain the organ of interest. Alternatively one could increase prediction density for organ sub-structures or smaller anatomical structures as needed. Our approach is therefore easily adapted to yield coarser bounding box detection results if desired. Additionally, our framework is easily extended to multi-organ segmentation. With the increasing availability of medical data, building a benchmark dataset now becomes essential to assess the generalization performance and scalability of high capacity machine learning models as used (Cuingnet et al., 2012; Glocker et al., 2012).

Some promising avenues for further exploration include the issue of how to regularize the ConvNet spatial prediction map by exploiting the three dimensions of the image volume without significant additional computational cost. The extension of the method to handle 3D inputs is also an avenue of exploration that is likely to yield improved results; however, one must deal with the variability of inter-slice spacing across different CT scans. The degree of human agreement can be an important factor for this type of segmentation. As such, our dataset would benefit from an additional ground truth labelling by a second expert.

6.6 Acknowledgements

This paper was supported in part by the Canada Research Chair in Medical Imaging and Assisted Interventions, the Natural Sciences and Engineering Research Council of Canada

and the MEDITIS training program. The authors would like to thank the developers of Theano (Bergstra et al., 2010; Bastien et al., 2012).

CHAPITRE 7 RÉSULTATS COMPLÉMENTAIRES

Le chapitre 6 a présenté des travaux portant sur la segmentation de reins par des réseaux à convolution. Le Tableau 7.1 montre une comparaison qualitative entre les différentes méthodes proposées dans la littérature.

Tableau 7.1 Tableau comparatif des différentes méthodes de segmentation pour les reins gauche et droit dans des images de CT selon le coefficient de Dice (DC).

Méthode	**# cas**	**Multiorgane**	**DC**	
			Droite	Gauche
Atlas hiérarchique (Wolz et al., 2012)	10	Oui	94	
Atlas + coupe de graphe (Chu et al., 2013)	100	Oui	90 ± 5	
Ligne de niveau (Khalifa et al., 2011)	21 Testé sur 14	Non	97 ± 0.02	
Forêts aléatoires et atlas* (Cuingnet et al., 2012)	233 Testé sur 179	Non	96 93 – 97	96 93 – 97
Forêts aléatoires (Glocker et al., 2012)	80 VC† avec 2 plis	Oui	62	68
ConvNet* (chapitre 6)	79 Testé sur 20	Non	95 92 – 95	93 90 – 94

* Les coefficients rapportés correspondent à la médiane, et aux 1^{er} et $3^{ème}$ quartiles. Autrement, il s'agit de la valeur moyenne, éventuellement accompagnée de l'écart-type
† Validation croisée.

Il n'existe pour l'instant aucun consensus au niveau des données ou des métriques utilisées pour mesurer la performance de segmentation. La métrique la plus couramment utilisée est le coefficient de Dice (voir Annexe A pour le détail de la formulation) qui mesure le recouvrement entre deux ensembles de points. Les objectifs diffèrent selon les études, il peut s'agir d'une segmentation de plusieurs organes de l'abdomen en même temps, ou de seulement un organe à la fois, comme proposé dans les travaux du chapitre 6. Les jeux de données diffèrent également, ce qui rend les comparaisons entre les différentes méthodes d'autant plus compliquées.

Toutefois, une analyse des coefficients de Dice à travers les études reste intéressante pour déterminer les tendances. Il semblerait que la construction d'un atlas aide au raffinement des segmentations pour obtenir de bons scores. Wolz et al. (2012) et Chu et al. (2013) obtiennent notamment des scores élevés mais leurs méthodes basées sur des atlas nécessitent plusieurs

heures pour obtenir les segmentations. L'une des principales différences entre les travaux de Cuingnet et al. (2012) et de Glocker et al. (2012) concerne l'utilisation d'atlas. Ces deux travaux utilisent des forêts aléatoires, mais Cuingnet et al. (2012) ont rajouté un atlas pour obtenir les segmentations, ce qui résulte en une meilleure performance. Khalifa et al. (2011) utilisent un a priori sur la forme des reins et leur jeu de données ne comporte pas que des sujets sains. Notre méthode, présentée au chapitre 6, permet d'arriver à des scores élevés, selon le coefficient de Dice, sans passer par l'utilisation d'atlas. De plus, la méthodologie permet une détection et une segmentation simultanée.

CHAPITRE 8 DISCUSSION GÉNÉRALE

8.1 Vers une nouvelle classification de la scoliose

Le chapitre 5 propose une nouvelle explication de la déformation chez des patients atteints de la scoliose idiopathique de l'adolescent. Ce regroupement a été réalisé grâce à des auto-encodeurs empilés, un algorithme d'apprentissage non supervisé, qui a permis d'apprendre une représentation latente des reconstructions de la colonne vertébrale des patients de la base de données. Les 663 patients inclus sont atteints de scoliose idiopathique de l'adolescent et résultent en 915 reconstructions qui couvrent tous les groupes de la classification Lenke. Onze différents groupes sont proposés. Chacun des groupes possède des caractéristiques qui leur sont propres.

L'article n'émet en aucun cas une nouvelle classification que les chirurgiens orthopédiques doivent suivre dans les mois ou années à venir. Il s'agit de la différence entre une invention et une innovation. Les inventions posent les fondamentaux et les jalons pour aboutir à une innovation qui sera adoptée par tous. Les travaux présentés au chapitre 5 sont dans la lignée des précédentes études réalisées précédemment (Sangole et al., 2009; Duong et al., 2006, 2009; Kadoury and Labelle, 2012). La différence majeure concerne l'inclusion de nombreux types de déformations, ce qui rend la base de données du chapitre 5 plus en lien avec les attentes des chirurgiens orthopédiques. L'ensemble de ces études contribue à la compréhensions des déformations de la colonne vertébrales chez les adolescents scoliotiques. L'objectif final est d'aider le chirurgien orthopédique à choisir la bonne correction à appliquer au niveau de la colonne vertébrale pour optimiser le résultat final.

Un travail sur la vulgarisation, l'explication, la compréhension et la reproductibilité de la méthodologie doit être réalisé pour conduire à une adoption de la part de la communauté médicale. Une classification a pour rôle de faciliter la communication entre praticiens hospitaliers lorsqu'il s'agit de décrire de manière simple les différentes déformations. À cela s'ajoutent des directives qui doivent être prises à plus haut niveau, par exemple au sein de la Scoliosis Research Society, pour standardiser l'évaluation de la scoliose afin de permettre des études multicentriques et un consensus entre les médecins pour décrire les déformations de la scoliose. Les travaux présentés au chapitre 5 vont d'ailleurs dans ce sens puisqu'ils rassemblent des patients provenant de neuf centres différents à travers l'Amérique du Nord.

8.2 Réseaux à convolution pour la détection et la segmentation d'organes dans des images médicales

Le chapitre 6 propose une méthode automatique de détection et de segmentation des reins par la classification de voxels avec un réseau à convolution. Cette méthodologie est réalisée en deux étapes. Dans un premier temps, le réseau à convolution est entrainé sur un ensemble de patches récoltés aléatoirement dans le jeu de données d'entrainement. Dans un second temps, l'architecture du réseau à convolution est transformée, sans altérer les paramètres appris, pour produire des segmentations de la dimension de l'image médicale.

Par ailleurs, la méthodologie présentée est facilement transférable et adaptable à une segmentation multiorgane à condition d'obtenir les données étiquetées au préalable. Il est d'ailleurs intéressant de mentionner que dans le cas de réseaux de neurones, avoir plus de classes tend à améliorer les résultats finaux (Ouyang et al., 2015). On peut alors s'attendre à une amélioration de la segmentation des reins si l'on rajoute d'autres organes à segmenter.

Dans la littérature récente, de nombreuses approches ont été proposées pour entrainer un ConvNet de bout en bout sans passer par une division en de multiples patches (Long et al., 2015). Malheureusement, l'imagerie médicale fait face à un inconvénient majeur qui empêche de transférer aisément ces méthodologies de segmentation d'images naturelles à des problèmes de segmentation d'images médicales. Si l'on prend l'exemple de la segmentation des reins, la région d'intérêt représente moins d'un pourcent du volume global de l'image. Cela veut dire que le volume de l'image est composé de 1% de voxels appartenant aux reins et de 99% de voxels appartenant à l'arrière-plan. Dans le cas où un entrainement du ConvNet se fait avec des images entières et non des patches, le ConvNet verra essentiellement – si ce n'est l'entière majorité du temps – des pixels appartenant à la classe de l'arrière-plan (c'est-à-dire le reste du corps humain). Même si la fonction objectif peut être pondérée pour compenser ce débalancement, la sensibilité de la classification des reins risque d'être très faible étant donné la trop faible proportion de voxels appartenant aux reins. En effet, un tel débalancement causerait un biais dans le ConvNet qui fait en sorte tendrait à classifier tous les pixels en tant qu'arrière-plan car il n'aurait pas appris de traits caractéristiques qui discriminent les reins des autres structures. Cet inconvénient peut néanmoins être résolu de plusieurs manières :

— Une méthode simple consisterait à réduire le champ de vision (FOV) du ConvNet. Au lieu de considérer l'image entière, un ConvNet spécialisé se concentrerait à prédire la classe des pixels au sein d'une boîte englobant les reins.
— Une seconde méthode plus fastidieuse consiste à étiqueter plus d'organes à l'intérieur

du corps humain. Dans ce cas-ci, le ConvNet apprendrait des traits caractéristiques qui sont plus discriminants grâce à la présence de nombreuses classes (Ouyang et al., 2015). La majorité des images qui ne contenaient qu'une seule classe auparavant (c'est-à-dire l'arrière-plan), contiennent désormais plus de classes (par exemple les poumons, le foie, le pancréas...).

De plus, les travaux du chapitre 6 peuvent conduire à d'autres heuristiques. La méthodologie en tant que telle permet une détection et une segmentation simultanée en un faible temps de calcul. La détection d'organes est notamment utile pour indexer les images médicales et pour définir une boîte englobant la région d'intérêt. Des méthodes en cascade à plusieurs échelles peuvent également raffiner à chaque étape la segmentation. L'idée est d'entrainer des réseaux à convolution qui sont de plus en plus spécialisés afin d'optimiser les résultats finaux. Les prédictions de la méthodologie actuelle restent contraintes à deux dimensions et n'exploitent pas la troisième dimension qui pourrait améliorer la régularisation de la segmentation.

8.3 Métriques et variabilité des données

Les métriques permettent de quantifier la performance d'une méthode. Le domaine biomédical fait face à une forte variabilité au sein des données, et les métriques utilisées ne parviennent souvent pas à capturer toute cette variabilité.

Prenons l'exemple de la classification de la scoliose. Comment chaque sous-groupe trouvé par la méthodologie proposée peut être caractérisé ? Au cours du chapitre 5 nous nous sommes référés aux indices géométriques traditionnels qui avaient été utilisés auparavant pour caractériser chacun des sous-groupes. Or la motivation des algorithmes d'apprentissage de représentations était de déroger à ces indices géométriques que les précédentes études utilisaient pour établir leurs sous-groupes. Même si notre utilisation des indices géométriques est différente – dans notre cas ils servent à expliquer les sous-groupes trouvés et non à former des sous-groupes comme précédemment – la quête portant sur la recherche des meilleurs indices géométriques reste un problème particulièrement ouvert. L'apport de l'expérience du chirurgien orthopédique est essentiel pour dépasser l'interprétation que l'on peut se faire juste en observant des métriques.

Le même constat s'applique pour la segmentation de reins. Les métriques utilisées dans le chapitre 6 sont détaillées en Annexe A. Un ensemble de métriques est nécessaire pour l'interprétation des résultats obtenus. Se fier seulement au coefficient de Dice n'est pas suffisant. En effet, il est impossible de savoir à partir de cette seule métrique si l'algorithme a sous- ou sur-segmenté les reins puisque seul le recouvrement est mesuré. Les mesures de distances,

que ce soit l'ASSD ou la distance de Hausdorff, permettent d'avoir une idée sur la forme de la segmentation mais demeurent sensibles aux points aberrants car elles privilégient les structures de forme arrondie et lisse. L'apport de l'expertise du médecin est alors crucial pour comprendre l'utilisation des segmentations dans son processus de décision. En concordance avec ces spécifications, une orientation pour favoriser une métrique en particulier peut être définie au sein de l'algorithme. Or, le chapitre 6 ne s'attèle qu'à l'exactitude de la classification. Si le coefficient de Dice est par exemple la métrique à favoriser, une nouvelle fonction objectif peut être dérivée pour que le réseau à convolution maximise cette métrique.

La présence d'une expertise médicale est indéniable nécessaire pour interpréter les résultats, patient par patient. En plus d'un forage des données pour comprendre les principaux facteurs de variation, des outils de visualisation sont également nécessaires pour faciliter l'interprétation des résultats. Ceci n'a pas fait l'objet de ce mémoire, mais les méthodes proposées ainsi que les résultats obtenus peuvent s'intégrer au sein d'un pipeline de visualisation. Un chirurgien orthopédique pourrait par exemple visualiser la représentation latente apprise par les auto-encodeurs empilés pour chacun des sous-groupes trouvés par les k-moyennes++. Un radiologue, ou un néphrologiste, pourrait par exemple visionner les cas où la segmentation du ConvNet est problématique, et estimer s'il est nécessaire d'y atteler plus de temps pour régler les défauts du ConvNet qui serait alors modifié pour répondre ce besoin. Au contraire, si l'erreur est futile, les ressources peuvent alors être allouées dans d'autres secteurs.

Par ailleurs, l'apprentissage automatique repose sur une hypothèse principale qui stipule que le jeu de données sur lequel le modèle va être entrainé est représentatif du problème que l'on cherche à résoudre. Dans le domaine biomédical, cela se traduit par des règles d'inclusion ou d'exclusion des patients lorsqu'il vient le temps de construire une base de données. Ces critères définis par le médecin sont impératifs pour réaliser des études rétrospectives ou prospectives qui répondent à leurs besoins et qui aient un impact significatif dans le domaine étudié. Il est en effet primordial de rassembler une cohorte qui soit la plus large possible pour satisfaire les enjeux liés au problème à résoudre.

Le médecin est donc au centre de toutes les méthodes proposées au cours de ce mémoire, car les informations additionnelles qui sont produites, soit par la classification des courbures de scolioses, soit par la classification de voxels dans une image médicale, jouent un rôle prépondérant dans son processus de décision.

8.3.1 Quel paradigme d'apprentissage ?

Les succès les plus marquants dans le domaine de l'apprentissage de représentations concernent l'apprentissage supervisé. De plus en plus de tâches reliées à la vision par ordinateur ont un

équivalent avec un réseau de neurones. On peut citer par exemple la prédiction de cartes de profondeur (Eigen et al., 2014), le flux optique (Weinzaepfel et al., 2013), ou encore la stéréoscopie (Fischer et al., 2015). Comme mentionné à la section 2.6.1, ceci est dû à une meilleure compréhension de l'entrainement des architectures profondes et de jeux de données étiquetés en grand nombre. Cette approche purement supervisée a sur-passé les résultats obtenus par des approches non-supervisées lors de la résurgence des algorithmes de représentations. Les approches non-supervisées ont en effet été utilisées pour initialiser des approches supervisées. Cette heuristique a aujourd'hui perdu de son intérêt au fil des années.

Au cours de ce mémoire, les algorithmes d'apprentissage non-supervisé et supervisé utilisés, à savoir les auto-encodeurs et les réseaux à convolution, partagent néanmoins de nombreuses composantes puisqu'ils sont tous deux des réseaux de neurones artificiels. Par exemple, l'initialisation des paramètres des auto-encodeurs débruitants et des ConvNets ont été réalisés selon la même heuristique, et la durée de l'entrainement a été déterminée par arrêt prématuré. Les méthodologies pour l'entrainement d'ANNs ne sont pas en vase clos, des avancées dans un domaine peuvent donc se transférer dans un autre et vice versa.

En plus de partager des composantes, les auto-encodeurs et réseaux à convolution peuvent être combinés pour former des auto-encodeurs à convolution. L'idée est de combiner la capacité des auto-encodeurs à apprendre de manière non-supervisée avec la capacité des réseaux à convolution de modéliser la topologie de l'image. Dans le cas de la scoliose, la méthodologie proposée au chapitre 5 repose sur la reconstruction en 3D de la colonne vertébrale à partir de rayons X. Cette reconstruction pourrait être réalisée à partir d'un auto-encodeur à convolution et servir en même temps à partitionner les patients en sous-groupes. Dans le cadre de la segmentation d'organes, un auto-encodeur à convolution pourrait par exemple apprendre des structures anatomiques de manière non-supervisée dans des images médicales. Un partitionnement peut alors s'avérer utile pour comprendre l'apprentissage qui a été réalisé.

Toutefois, l'apprentissage non-supervisé demeure primordial dans le domaine biomédical où il est difficile d'avoir accès à des annotations réalisées par des experts. Ce coût en terme de temps et d'argent pour obtenir des données étiquetées laisse penser qu'un apprentissage semi-supervisé pourrait être favorisé à l'avenir. Si un hôpital est en capacité de fournir des milliers d'images médicales, est-il réaliste de demander à des experts d'étiqueter une à une chacune des images ? La solution idéale s'orienterait vers une situation où l'expert ne produit des étiquettes que pour les cas qu'il estime importants, ou qui sont à fort risque d'engendrer des erreurs. L'algorithme infèrerait alors à partir de ce nombre limité d'exemples étiquetés les représentations nécessaires pour résoudre la tâche spécifiée. Un exemple similaire peut également s'appliquer avec le cas de la scoliose idiopathique de l'adolescent où le chirurgien

imposerait des contraintes à l'algorithme. Une information a priori pourrait par exemple spécifier que certains patients doivent être regroupés ensemble alors d'autres doivent être absolument séparés. Encore une fois, à partir de ce faible nombre de données étiquetées, l'algorithme infèrerait les principaux facteurs de variation qui discriminent les différents sous-groupes.

CHAPITRE 9 CONCLUSION

Les travaux réalisés au cours de ce mémoire montrent l'intérêt de l'apprentissage de représentations pour la classification d'images biomédicales pour des projets ayant une portée clinique concrète.

Dans un premier temps, des auto-encodeurs empilés ont servi à apprendre une représentation latente des reconstructions de la colonne vertébrale de patients atteints de la scoliose idiopathique de l'adolescent. Onze sous-groupes statistiquement significatifs ont été trouvés pour expliquer les déformations de la colonne vertébrale au sein de la base de données qui rassemble des patients de tous les types de la classification Lenke. Le cadre métrologique étant posé, une analyse clinique plus poussée reste à réaliser pour mieux comprendre ce qui caractérise chacun de ces sous-groupes.

Dans un second temps, un réseau à convolution a été utilisé pour segmenter les reins dans des images de tomodensitométrie avec agent de contraste chez des patients qui présentent de fortes complications rénales. La performance réalisée demeure élevée avec un coefficient de Dice de 94,35% pour le rein gauche et 93,07% pour le rein droit, et ce avec un temps de calcul réduit. Une comparaison plus poussée reste néanmoins à réaliser vis-à-vis des algorithmes utilisés dans la littérature pour la segmentation de reins.

Les algorithmes d'apprentissage de représentations sont en constante évolution aujourd'hui, et s'améliorent à un rythme effréné. Les opportunités dans le domaine biomédical sont donc grandissantes à condition de rassembler une base de données qui soit représentative et d'une taille suffisante. En effet, ces méthodes d'apprentissage de représentations reposent sur la disponibilité de nombreuses données, qui se doivent d'être représentatives du problème que l'on souhaite résoudre. Le transfert de technologie vers le domaine biomédical prend donc plus de temps car une expertise médicale pointue est nécessaire à la fois pour rassembler des observations pour une base de données, mais aussi pour interpréter les résultats obtenus.

RÉFÉRENCES

D. Arthur et S. Vassilvitskii, “k-means++ : The advantages of careful seeding”, dans *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

M. A. Asher et D. C. Burton, “Adolescent idiopathic scoliosis : natural history and long term treatment effects”, *Scoliosis*, vol. 1, no. 1, p. 2, 2006.

P. Baqué et B. Maes, *Manuel pratique d’anatomie : descriptive, topographique, fonctionnelle, clinique et embryologique*. Ellipses, 2008.

F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, et Y. Bengio, “Theano : new features and speed improvements”, Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

Y. Bengio, A. Courville, et P. Vincent, “Representation learning : A review and new perspectives.” *IEEE transactions on pattern analysis and machine intelligence*, 2013.

Y. Bengio, “Practical recommendations for gradient-based training of deep architectures”, dans *Neural Networks : Tricks of the Trade*. Springer, 2012, pp. 437–478.

Y. Bengio, I. J. Goodfellow, et A. Courville, “Deep learning”, 2015, book in preparation for MIT Press. En ligne : `http://www.iro.umontreal.ca/~bengioy/dlbook`

J. Bergstra et Y. Bengio, “Random search for hyper-parameter optimization”, *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.

J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, et Y. Bengio, “Theano : a CPU and GPU math expression compiler”, dans *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

C. Chu, M. Oda, T. Kitasaka, K. Misawa, M. Fujiwara, Y. Hayashi, Y. Nimura, D. Rueckert, et K. Mori, “Multi-organ segmentation based on spatially-divided probabilistic atlas from 3d abdominal ct images”, dans *MICCAI*. Springer, 2013, pp. 165–172.

D. Ciresan, A. Giusti, L. M. Gambardella, et J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images”, dans *NIPS*, 2012, pp. 2843–2851.

A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, et K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans", *Medical image analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, et R. Ardon, "Automatic detection and segmentation of kidneys in 3d ct images using random forests", dans *MICCAI*. Springer, 2012, pp. 66–74.

J. Dubousset, G. Charpak, I. Dorion, W. Skalli, F. Lavaste, J. Deguise, G. Kalifa, et S. Ferey, "Une nouvelle imagerie ostéo-articulaire basse dose en position debout : le système eos", *Radioprotection*, vol. 40, no. 02, pp. 245–255, 2005.

L. Duong, F. Cheriet, et H. Labelle, "Three-dimensional classification of spinal deformities using fuzzy clustering", *Spine*, vol. 31, no. 8, pp. 923–930, 2006.

L. Duong, J.-M. Mac-Thiong, F. Cheriet, et H. Labelle, "Three-dimensional subclassification of lenke type 1 scoliotic curves", *Journal of spinal disorders & techniques*, vol. 22, no. 2, pp. 135–143, 2009.

D. Eigen, C. Puhrsch, et R. Fergus, "Depth map prediction from a single image using a multi-scale deep network", dans *NIPS*, 2014, pp. 2366–2374.

EOS Espace média, "Système eos". En ligne : `http://www.eos-imaging.com/fr/news-events/pressroom.html`

H. Fanet, *Imagerie médicale à base de photons : Radiologie, tomographie X, tomographie gamma et positons, imagerie optique*, série Traité EGEM. : Série électronique et microélectronique. Lavoisier, 2010.

P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, et T. Brox, "Flownet : Learning optical flow with convolutional networks", *CoRR*, vol. abs/1504.06852, 2015.

M. Freiman, A. Kronman, S. J. Esses, L. Joskowicz, et J. Sosna, "Non-parametric iterative model constraint graph min-cut for automatic kidney segmentation", dans *MICCAI*. Springer, 2010, pp. 73–80.

A. Giusti, D. C. Ciresan, J. Masci, L. M. Gambardella, et J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks", *CoRR*, vol. abs/1302.1700, 2013.

B. Glocker, O. Pauly, E. Konukoglu, et A. Criminisi, “Joint classification-regression forests for spatially structured multi-object segmentation”, dans *ECCV*. Springer, 2012, pp. 870–881.

X. Glorot et Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks”, dans *AISTATS*, 2010, pp. 249–256.

X. Glorot, A. Bordes, et Y. Bengio, “Deep sparse rectifier neural networks”, dans *AISTATS*, 2011, pp. 315–323.

E. Godaux, *Les neurones, les synapses et les fibres musculaires*. Masson, 1994.

G. Godet, M.-H. Fleron, E. Vicaut, A. Zubicki, M. Bertrand, B. Riou, E. Kieffer, et P. Coriat, “Risk factors for acute postoperative renal failure in thoracic or thoracoabdominal aortic surgery : a prospective study”, *Anesthesia & Analgesia*, vol. 85, no. 6, pp. 1227–1232, 1997.

L. R. Goodman, “The beatles, the nobel prize, and ct scanning of the chest”, *Radiologic Clinics of North America*, vol. 48, no. 1, pp. 1–7, 2010.

P. Grangeat, *La Tomographie médicale : Imagerie morphologique et imagerie fonctionnelle*, série IC2 : Série Traitement du signal et de l’image. Lavoisier, 2002.

Z. Hall, *Introduction à la neurobiologie moléculaire*. Médecine Sciences Publications, 1994.

L. K. Hansen et P. Salamon, “Neural network ensembles”, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 993–1001, 1990.

G. Hinton et R. Salakhutdinov, “Reducing the dimensionality of data with neural networks.” *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 504–507, 2006. DOI : 10.1126/science.1127647

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups”, *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

T. Illés et S. Somoskeöy, “The eos$^{\text{TM}}$ imaging system and its uses in daily orthopaedic practice”, *International orthopaedics*, vol. 36, no. 7, pp. 1325–1331, 2012.

N. Jones, “Computer science : The learning machines.” *Nature*, vol. 505, no. 7482, p. 146—148, 2014.

S. Kadoury et H. Labelle, “Classification of three-dimensional thoracic deformities in adolescent idiopathic scoliosis from a multivariate analysis”, *European Spine Journal*, vol. 21, no. 1, pp. 40–49, 2012.

S. Kadoury, F. Cheriet, et H. Labelle, “Personalized x-ray 3-d reconstruction of the scoliotic spine from hybrid statistical and image-based models”, *Medical Imaging, IEEE Transactions on*, vol. 28, no. 9, pp. 1422–1435, 2009.

S. Kadoury, J. Shen, et S. Parent, “Global geometric torsion estimation in adolescent idiopathic scoliosis”, *Medical & biological engineering & computing*, vol. 52, no. 4, pp. 309–319, 2014.

W. A. Kalender, “X-ray computed tomography”, *Physics in medicine and biology*, vol. 51, no. 13, p. R29, 2006.

F. Khalifa, A. Elnakib, G. M. Beache, G. Gimel’farb, M. A. El-Ghar, R. Ouseph, G. Sokhadze, S. Manning, P. McClure, et A. El-Baz, “3d kidney segmentation from ct images using a level set approach guided by a novel stochastic speed function”, dans *MICCAI*. Springer, 2011, pp. 587–594.

H. A. King, J. H. Moe, D. S. Bradford, et R. B. Winter, “The selection of fusion levels in thoracic idiopathic scoliosis.” *The Journal of Bone & Joint Surgery*, vol. 65, no. 9, pp. 1302–1313, 1983.

A. Krizhevsky, I. Sutskever, et G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, dans *NIPS*, 2012, pp. 1097–1105.

H. Labelle, C.-E. Aubin, R. Jackson, L. Lenke, P. Newton, et S. Parent, “Seeing the spine in 3d : how will it change what we do ?” *Journal of Pediatric Orthopaedics*, vol. 31, pp. S37–S45, 2011.

Y. LeCun, L. Bottou, Y. Bengio, et P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998. DOI : 10.1109/5.726791

Y. A. LeCun, L. Bottou, G. B. Orr, et K.-R. Müller, “Efficient backprop”, dans *Neural networks : Tricks of the trade*. Springer, 2012, pp. 9–48.

J. Legaye, G. Duval-Beaupere, J. Hecquet, et C. Marty, “Pelvic incidence : a fundamental pelvic parameter for three-dimensional regulation of spinal sagittal curves”, *European Spine Journal*, vol. 7, no. 2, pp. 99–103, 1998.

L. G. Lenke, R. R. Betz, J. Harms, K. H. Bridwell, D. H. Clements, T. G. Lowe, et K. Blanke, "Adolescent idiopathic scoliosis", *The Journal of Bone & Joint Surgery*, vol. 83, no. 8, pp. 1169–1181, 2001.

J. Long, E. Shelhamer, et T. Darrell, "Fully convolutional networks for semantic segmentation", dans *CVPR*, 2015, pp. 3431–3440.

B. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)", *IEEE Transactions on Medical Imaging*, p. 33, 2014.

A. Y. Ng, "Feature selection, l 1 vs. l 2 regularization, and rotational invariance", dans *ICML*. ACM, 2004, p. 78.

W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy *et al.*, "Deepid-net : Deformable deep convolutional neural networks for object detection", dans *CVPR*, 2015, pp. 2403–2412.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn : Machine learning in python", *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

V. Pomero, D. Mitton, S. Laporte, J. A. de Guise, et W. Skalli, "Fast accurate stereoradiographic 3d-reconstruction of the spine using a combined geometric and statistic model", *Clinical Biomechanics*, vol. 19, no. 3, pp. 240–247, 2004.

P. Poncet, J. Dansereau, et H. Labelle, "Geometric torsion in idiopathic scoliosis : three-dimensional analysis and proposal for a new classification", *Spine*, vol. 26, no. 20, pp. 2235–2243, 2001.

I. Ponseti et B. Friedman, "Prognosis in idiopathic scoliosis." *The Journal of bone and joint surgery. American volume*, vol. 32, no. 2, p. 381, 1950.

A. Ramé et S. Thérond, *Anatomie et physiologie : Aide-soignant et auxiliaire de puériculture*. Elsevier Health Sciences France, 2012.

S. Ray et R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation", dans *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*. India, 1999, pp. 137–143.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge", *International Journal of Computer Vision*, pp. 1–42, 2014.

A. P. Sangole, C.-E. Aubin, H. Labelle, I. A. Stokes, L. G. Lenke, R. Jackson, et P. Newton, "Three-dimensional classification of thoracic scoliotic curves", *Spine*, vol. 34, no. 1, pp. 91–99, 2009.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, et Y. LeCun, "Overfeat : Integrated recognition, localization and detection using convolutional networks", *CoRR*, vol. abs/1312.6229, 2013.

K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *CoRR*, vol. abs/1409.1556, 2014.

J. Snoek, H. Larochelle, et R. P. Adams, "Practical bayesian optimization of machine learning algorithms", dans *NIPS*, 2012, pp. 2951–2959.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, et R. Salakhutdinov, "Dropout : A simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

I. Stokes, L. Bigalow, et M. Moreland, "Measurement of axial rotation of vertebrae in scoliosis." *Spine*, vol. 11, no. 3, p. 213, 1986.

I. A. Stokes, "Three-dimensional terminology of spinal deformity : A report presented to the scoliosis research society by the scoliosis research society working group on 3-d terminology of spinal deformity." *Spine*, vol. 19, no. 2, pp. 236–248, 1994.

I. A. Stokes, A. P. Sangole, et C.-E. Aubin, "Classification of scoliosis deformity 3-d spinal shape by cluster analysis", *Spine*, vol. 34, no. 6, pp. 584–590, 2009.

I. Sutskever, J. Martens, G. Dahl, et G. Hinton, "On the importance of initialization and momentum in deep learning", dans *ICML*, 2013, pp. 1139–1147.

L. J. van der Maaten, E. O. Postma, et H. J. van den Herik, "Dimensionality reduction : A comparative review", *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.

P. Vincent, H. Larochelle, Y. Bengio, et P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders", dans *ICML*. ACM, 2008, pp. 1096–1103.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, et P.-A. Manzagol, “Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion”, *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

S. Waldt, A. Gersing, et M. Brügel, “Measurements and classifications in spine imaging.” dans *Seminars in musculoskeletal radiology*, vol. 18, no. 3, 2014, pp. 219–227.

S. L. Weinstein, L. A. Dolan, J. C. Cheng, A. Danielsson, et J. A. Morcuende, “Adolescent idiopathic scoliosis”, *The Lancet*, vol. 371, no. 9623, pp. 1527–1537, 2008.

P. Weinzaepfel, J. Revaud, Z. Harchaoui, et C. Schmid, “Deepflow : Large displacement optical flow with deep matching”, dans *ICCV*. IEEE, 2013, pp. 1385–1392.

Wikimedia Commons. (2007) Schéma d’un neurone, commenté en français. File :Neurone - commenté.svg. En ligne : `https://commons.wikimedia.org/wiki/File:Neurone_-_commenté.svg`

——. (2010) La colonne vertébrale - anatomie. File :Spine Anatomy Kisco.JPG. En ligne : `https://commons.wikimedia.org/wiki/File:Spine_Anatomy_Kisco.JPG`

Wikiversity Journal of Medicine. (2014) Blausen gallery 2014. En ligne : `DOI: 10.15347/wjm/2014.010`

R. Wolz, C. Chu, K. Misawa, K. Mori, et D. Rueckert, “Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases”, dans *MICCAI*. Springer, 2012, pp. 10–17.

M. Wybier et P. Bossard, “Musculoskeletal imaging in progress : the eos imaging system”, *Joint Bone Spine*, vol. 80, no. 3, pp. 238–243, 2013.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes *et al.*, “The human splicing code reveals new insights into the genetic determinants of disease”, *Science*, vol. 347, no. 6218, p. 1254806, 2015.

ANNEXE A MÉTRIQUES QUANTITATIVES POUR LA SEGMENTATION

Le coefficient de Dice (DC) mesure la similarité entre deux volumes :

$$DC(A,B) = 2\frac{|A \cap B|}{|A| + |B|} \tag{A.1}$$

où A et B sont deux volumes différents de voxels.

La distance moyenne symétrique surfacique (ASSD) dénote la distance moyenne entre les points de la surface des deux volumes dans les deux directions. La distance moyenne surfacique (ASD) se calcule par :

$$ASD(A,B) = \frac{\sum_{a \in A} \min_{b \in B} d(a,b)}{|A|} \tag{A.2}$$

où A et B sont deux ensembles de points, et $d(A,B)$ est la distance Euclidienne entre les points a and b. Pour obtenir la symétrie :

$$ASSD(A,B) = \frac{ASD(A,B) + ASD(B,A)}{2} \tag{A.3}$$

La distance d'Hausdorff dénote la distance maximale entre deux surfaces de points d'un volume :

$$HD(A,B) = \max\{\max_{a \in A} \min_{b \in B} d(a,b), \max_{b \in B} \min_{a \in A} d(b,a)\} \tag{A.4}$$